5-2011

PRETICTIVE BIOINFORMATIC METHODS FOR ANALYZING GENES AND PROTEINS

Shaolei Teng
Clemson University, steng@clemson.edu

PRETICTIVE BIOINFORMATIC METHODS FOR ANALYZING GENES AND PROTEINS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Biochemistry and Molecular Biology

---

by
Shaolei Teng
May 2011

Accepted by:
Dr. Liangjiang Wang, Committee Chair
Dr. Emil Alexov
Dr. Charles Schwartz
Dr. Chin-Fu Chen

ABSTRACT

Since large amounts of biological data are generated using various high-throughput technologies, efficient computational methods are important for understanding the biological meanings behind the complex data. Machine learning is particularly appealing for biological knowledge discovery. Tissue-specific gene expression and protein sumoylation play essential roles in the cell and are implicated in many human diseases. Protein destabilization is a common mechanism by which mutations cause human diseases. In this study, machine learning approaches were developed for predicting human tissue-specific genes, protein sumoylation sites and protein stability changes upon single amino acid substitutions. Relevant biological features were selected for input vector encoding, and machine learning algorithms, including Random Forests and Support Vector Machines, were used for classifier construction. The results suggest that the approaches give rise to more accurate predictions than previous studies and can provide valuable information for further experimental studies. Moreover, seeSUMO and MuStab web servers were developed to make the classifiers accessible to the biological research community.

Structure-based methods can be used to predict the effects of amino acid substitutions on protein function and stability. The nonsynonymous Single Nucleotide Polymorphisms (nsSNPs) located at the protein binding interface have dramatic effects on protein-protein interactions. To model the effects, the nsSNPs at the interfaces of 264 protein-protein complexes were mapped on the protein structures using homology-based methods. The results suggest that disease-causing nsSNPs tend to destabilize the

electrostatic component of the binding energy and nsSNPs at conserved positions have significant effects on binding energy changes. The structure-based approach was developed to quantitatively assess the effects of amino acid substitutions on protein stability and protein-protein interaction. It was shown that the structure-based analysis could help elucidate the mechanisms by which mutations cause human genetic disorders. These new bioinformatic methods can be used to analyze some interesting genes and proteins for human genetic research and improve our understanding of their molecular mechanisms underlying human diseases.

DEDICATION

I dedicate my dissertation to my parents, wife and two daughters. Without their love, the completion of this work would not have been possible.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Liangjiang Wang, for all of his guidance, patience and financial support. I am grateful to him for his expertise in machine learning and help in scientific writing. I appreciate the help of other committee members: Dr. Emil Alexov for his advice and guidance on my structure modeling studies, Dr. Charles Schwartz for his academic advice on my human genetic studies, Dr. Chin-Fu Chen for teaching me bioinformatic knowledge. I also want to express gratitude to all faculty, staff and students in Department of Genetics and Biochemistry at Clemson University and Greenwood Genetic Center for providing help for my PhD studies.

TABLE OF CONTENTS

Page

TITLE PAGE i

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGMENTS v

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

I. INTRODUCTION 1

II. A MACHINE LEARNING APPROACH FOR PREDICTING HUMAN TISSUE-SPECIFIC GENES USING MICROARRAY EXPRESSION DATA 11

Abstract 11
Background 12
Methods 15
Results and Discussion 21
Conclusion 34
References 35

III. PREDICTING PROTEIN SUMOYLATION SITES FROM SEQUENCE FEATURES 39

Abstract 39
Background 40
Methods 42
Results and Discussion 49
Conclusion 61
References 62

IV. SEQUENCE FEATURE-BASED PREDICTION OF PROTEIN STABILITY CHANGES UPON AMINO ACID SUBSTITUTIONS 65

Abstract 65
Background 66
Methods 68
Results and Discussion 74
Conclusion 85
References 86

V. MODELING EFFECTS OF HUMAN SINGLE NUCLEOTIDE POLYMORPHISMS ON PROTEIN-PROTEIN INTERACTIONS 89

Abstract 89
Introduction 90
Methods 94
Results and Discussion 100
Conclusion 115
References 119

VI. STRUCTURAL ASSESSMENT OF THE EFFECTS OF AMINO ACID SUBSTITUTIONS ON PROTEIN STABILITY AND PROTEIN-PROTEIN INTERACTION 126

Abstract 126
Introduction 126
Methods 130
Results and Discussion 137
Conclusions 146
References 147

VII. CONCLUSIONS 152

APPENDICES 154

A: Publications resulting from the present research 155
B: Additional files for predicting human tissue-specific genes 157
C: Additional files for predicting protein sumoylation sites 158
D: Additional data regarding the effect of nsSNPs on binding energy with respect to amino acid characteristics 172

LIST OF TABLES

Table Page

2.1 Comparison of Random Forest and Support Vector Machine classifiers for predicting tissue-specific genes 24

2.2 GO term enrichment analysis of predicted brain-specific genes 26

2.3 List of high-scoring genes with specific expression in the brain..... 29

2.4 GO term enrichment analysis of predicted liver-specific genes 31

2.5 List of high-scoring genes with specific expression in the liver..... 32

2.6 Random Forest classifiers for predicting tissue-selective genes 34

3.1 Effect of sequence context on predictive performance of Random Forest classifiers 51

3.2 Comparison of Random Forest and Support Vector Machine classifiers constructed with ΨKXE+/-2 (w = 8). 55

3.3 Effect of evolutionary information on protein sumoylation site prediction..... 56

3.4 Comparison of classifier performance using an independent test dataset..... 59

4.1 Effect of window sizes on sequence-based prediction of protein stability changes..... 75

4.2 Predictive performance of classifiers constructed using single sequence features..... 77

4.3 Predictive performance of classifiers constructed by combining the best single features 81

4.4 Predictive performance of classifiers constructed using the optimal subsets of sequence features 82

List of Tables (Continued)

Table Page

5.1 Parameters of distributions of total binding energy difference and its components together with the corresponding P-values.....101

6.1 The effects of five amino acid substitutions on protein stability.....138

6.2 The effects of five amino acid substitutions on protein-protein interaction 141

C.1 The list of 457 experimentally verified sumoylation sites in 263 proteins.....158

C.2 List of 40 biological features used in chapter three.....170

D.1 Parameters of distribution of total binding energy difference and its components 173

LIST OF FIGURES

Figure Page

2.1 Schematic diagram of the approach for predicting tissue-specific genes 16

2.2 Visualization of known tissue-specific gene expression patterns..... 22

2.3 ROC curves to compare the performances of Random Forest (RF) and Support Vector Machine (SVM) classifiers for predicting tissue-specific genes..... 25

2.4 Visualization of predicted tissue-specific gene expression patterns 27

3.1 The sequence logo of the protein sumoylation motif (ΨKXE) and its flanking residues..... 50

3.2 ROC curves to show the effect of context information for sumoylation site prediction..... 52

3.3 Performance comparisons of Random Forest (RF) and Support Vector Machine (SVM) classifiers..... 54

3.4 ROC curves to show the effect of evolutionary information on classifier performance..... 57

3.5 Sample output from the seeSUMO web server 61

4.1 ROC curves to show the effect of context information on prediction of protein stability changes upon amino acid substitutions..... 76

4.2 ROC curves to show the different performance levels of classifiers constructed using individual sequence features..... 78

4.3 ROC curves for sequence-based prediction of protein stability changes using multiple sequence features..... 83

4.4 Sample output from the MuStab web server..... 85

List of Figures (Continued)

Figure Page

5.1 Distribution of $\Delta\Delta\Delta G_{el}(nsSNP)$ and $\Delta\Delta\Delta G_{tot}(nsSNP)$ for OMIM and non-OMIM cases 102

5.2 Illustration of nsSNPs at interface of protein-protein complexes..... 104

5.3 Multiple sequence alignment (MSA) 111

5.4 The change of the binding energy as a function of the amino acid conservation 113

5.5 The change of the binding energy as a function of calculated proton uptake/release..... 114

6.1 Schematic diagram of the approach for assessing the effects of amino acid substitutions on protein stability and protein-protein interaction..... 131

6.2 Illustration of two amino acid substitutions (A111V and T119N) in human HBA2..... 141

6.3 Illustration of two disease-causing mutations (Q61K and A146T) in human HRAS..... 144

6.4 Illustration of the A693V mutation in human ZBTB20..... 146

D-1 Distribution of $\Delta\Delta\Delta G_{tot}(nsSNP)$, $\Delta\Delta\Delta G_{el}(nsSNP)$ and $\Delta\Delta\Delta G_{vdw}(nsSNP)$ in respect with physico-chemical properties of amino acids 174

CHAPTER ONE

INTRODUCTION

With the development of various genome sequencing projects and high-throughput technologies, modern biology has entered into a “data explosion” era. However, such large amounts of biological data bring the so-called "data rich, information poor" problem. On one hand, more and more biological data are generated by experimental studies, ranging from genomics to proteomics. On the other hand, it is not easy to extract useful information from the biological data, and the underlying molecular mechanisms remain elusive. There is a significant need for developing efficient computational methods to discover biological knowledge hidden in the massive and heterogeneous datasets.

Machine learning is a broad research field with wide applications in business, engineering and science. It focuses on designing and developing computer algorithms to improve predictive performance based on training data instances. Machine learning approaches such as Bayesian Networks, Hidden Markov Models, Neural Networks and Genetic Programming have been applied to various scientific fields including natural language processing, computer vision, search engine development, medical diagnosis and bioinformatics [1]. Machine learning can be used to recognize hidden patterns in data, and thus is particularly appealing for biological knowledge discovery in bioinformatic studies. There are different types of learning including unsupervised learning and supervised learning. Unsupervised learning can discover unknown clusters or detect anomalies from unlabeled data instances. It has been used to analyze genes associated

with human diseases. For example, clustering methods have been applied to analyses of gene expression data from different cancer samples including breast tumour samples [2], prostate cancer samples [3] and colon cancer samples [4]. By contrast, the training data instances used for supervised learning are labeled with the known information. Supervised learning can recognize hidden patterns in the labeled examples, and the resulting model can be used to make predictions for new data instances. It has been utilized for analyzing some important protein functions, such as protein secondary structures [5], functional residues [6], protein stability [7] and protein-protein interaction networks [8].

Supervised machine learning algorithms such as Support Vector Machines (SVMs) and Random Forests (RFs) have found wide applications in bioinformatic studies. SVMs can transform the training data into a feature space using kernel functions and then separate the data by a maximum-margin hyperplane [9]. SVMs have been used for predicting DNA/RNA-binding residues [10, 11], protein-protein interaction [12], subcellular localization [13] and protein stability changes upon mutations [14]. RFs are ensemble learning algorithms which can handle a large number of input variables and avoid model overfitting. It combines the votes made by the independent decision trees, and gives the most popular class as the output. RFs are becoming popular in various bioinformatic fields including structure classification [15], protein interaction site prediction [16], DNA-binding residues identification [17] and drug sensitivity prediction [18].

Tissue-specific gene plays a key role in the pathogenesis of many human diseases [19]. Several statistical approaches, including Akaike's information criterion [20], Shannon entropy [6] and hypothesis testing [21], have been utilized for identifying the tissue-specific genes using microarray expression data. However, the statistical methods assign an equal weight to each observation and do not use biological knowledge for predictions. A SVM-based approach has been developed to predict tissue-specific genes in *Caenorhabditis elegans* [22], but whether the machine learning methods can be used for predicting human tissue-specific genes is still unknown. Protein sumoylation is important for many cellular processes and any alterations in the process may cause various human diseases. Several computational methods such as SUMOpre [23], SUMOsp [24] and SUMOsp2.0 [25] have been developed for predicting sumoylation sites. To understand how a single amino acid substitution changes protein stability can help elucidate the molecular mechanism by which the mutations cause human diseases. Machine learning approaches such as I-Mutant2.0 [14] and iPTREE-STAB [26] have recently been applied to sequence-based prediction of protein stability changes upon mutations. However, little domain-specific knowledge in terms of relevant biological features was used in the studies for analyzing protein functions. In this study, machine learning approaches were developed for predicting human tissue-specific genes (chapter two), protein sumoylation sites (chapter three) and protein stability changes upon single amino acid substitutions (chapter four).

Supervised learning algorithms, including RFs and SVMs, have new applications in the present study. RFs and SVMs were used to identify human tissue-specific genes

with microarray gene expression data and predict protein sumoylation sites from protein sequence information. RF classifiers were found to outperform SVM models, and the approaches can provide useful information for further experimental studies. Furthermore, SVMs were applied to sequence-based prediction of protein stability changes upon amino acid substitutions. The supervised learning algorithms were used to develop seeSUMO (http://bioinfo.ggc.org/seesumo) and MuStab (http://bioinfo.ggc.org/mustab) web servers for predicting protein sumoylation sites and protein stability changes upon single amino acid substitutions, respectively.

The novelty of our approach is that the biological knowledge was used for classifier construction. Relevant biological features can be selected to construct classifiers and improve the predictive performance of classifiers. For example, the use of biochemical features and evolutionary information for input vector encoding can significantly improve the predictive performance of DNA-binding site prediction [11]. In this study, relevant features representing biological knowledge were used to encode input variables for sequence-based predictions of protein sumoylation sites and protein stability changes upon single amino acid substitutions. It was shown that the use of relevant biological features for classifier construction can significantly improve the predictive performances of classifiers.

The improvement of experimental determination of protein 3D structures and computational modeling [27, 28] made it possible to predict the effects of mutations by mapping them on the corresponding structures or models. Protein structural information has been used in many studies to reveal the role of amino acid substitutions on protein

function and stability. Previous studies on human non-synonymous Single Nucleotide Polymorphisms (nsSNP) in disease candidate genes revealed that approximately 70% of the deleterious mutations are located in the structurally and/or functionally important sites [29-32]. A structure-based approach that models residue-residue interaction networks was developed recently [33], and graph theoretical measures were used to predict the residues that are important for structural stability. The results suggest that mutations impact protein function and stability by affecting their structures, which in turn may cause changes in protein-protein interactions.

It has been estimated that each person may have 24,000 - 40,000 nsSNPs, and there are a total of 67,000 - 200,000 common nsSNPs in the human population [34]. Previous study suggest that approximately 25% of nsSNPs in the human population might be deleterious to protein function [35], and 88% of disease-associated nsSNPs are located in the voids/pockets important for protein-protein interactions [32]. The nsSNPs located at the binding interface or active site cleft may cause a series of changes, such as disruption of salt bridges, breakage of hydrogen bonds and alteration of binding affinity. In chapter five, we investigated the effects of nsSNPs at the interfaces of 264 protein-protein complexes using a structure-based method. The nsSNPs were mapped on the protein structures and their effects on the binding energy were investigated with CHARMM force field and continuum electrostatic calculation. The findings reveal that disease-causing nsSNPs tend to destabilize the electrostatic component of the binding energy on protein-protein interactions and nsSNPs at conserved positions can lead to a large variance of the binding energy changes.

The sequence/structure-based computational methods developed in this study can be used to analyze proteins, in which mutations may cause human genetic disorders such as intellectual disability. Intellectual disability is the most frequent developmental disability with an estimated incidence of 1-3% of people worldwide. It is often caused by loss-of-function mutations in associated genes. For example, several deleterious mutations in the spermine synthase (SMS) gene were found to cause Snyder-Robinson syndrome, an X-linked recessive disease with mild-to-moderate intellectual disability [36]. Recently, a structure-based approached was utilized to predict the effects of three missense SMS mutations causing Snyder-Robinson syndrome on protein stability, flexibility and interactions [37]. In chapter six, the structure-based approach is described for quantitatively assessing the effects of amino acids on protein stability and protein-protein interaction using homology modeling and free energy calculation methods. The results suggest that the structure-based approach together with sequence-based methods can provide useful information for characterizing mutations associated with intellectual disability in human genetic studies and elucidating the molecular mechanisms by which the mutations cause intellectual disability.

REFERENCES

1. Mjolsness E, DeCoste D: Machine learning for science: state of the art and future prospects. *Science* 2001, 293(5537):2051-2055.

2. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: Molecular portraits of human breast tumours. *Nature* 2000, 406(6797):747-752.

3. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001, 412(6849):822-826.

4. Getz G, Gal H, Kela I, Notterman DA, Domany E: Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics* 2003, 19(9):1079-1089.

5. Babaei S, Geranmayeh A, Seyyedsalehi SA: Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. *Comput Methods Programs Biomed* 2010, 100(3):237-247.

6. Vacic V, Iakoucheva LM, Lonardi S, Radivojac P: Graphlet kernels for prediction of functional residues in protein structures. *J Comput Biol* 2010, 17(1):55-72.

7. Capriotti E, Fariselli P, Casadio R: A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* 2004, 20 Suppl 1:i63-68.

8. Yamanishi Y, Vert JP, Kanehisa M: Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 2004, 20 Suppl 1:i363-370.

9. Noble WS: What is a support vector machine? *Nat Biotechnol* 2006, 24(12):1565-1567.

10. Wang L, Brown SJ: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006, 34(Web Server issue):W243-248.

11. Wang L, Huang C, Yang MQ, Yang JY: BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010, 4 Suppl 1:S3.

12. Gonzalez AJ, Liao L: Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines. *BMC Bioinformatics* 2010, 11:537.

13. Mak MW, Guo J, Kung SY: PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM. *IEEE/ACM Trans Comput Biol Bioinform* 2008, 5(3):416-422.

14. Capriotti E, Fariselli P, Casadio R: I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 2005, 33(Web Server issue):W306-310.

15. Jain P, Hirst JD: Automatic structure classification of small proteins using random forest. *BMC Bioinformatics* 2010, 11:364.

16. Chen XW, Jeong JC: Sequence-based Prediction of Protein Interaction Sites with an Integrative Method. *Bioinformatics* 2009.

17. Wang L, Yang MQ, Yang JY: Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* 2009, 10 Suppl 1:S1.

18. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, Fine HA: Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* 2011, 27(2):220-224.

19. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S: A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 2008, 105(52):20870-20875.

20. Kadota K, Nishimura S, Bono H, Nakamura S, Hayashizaki Y, Okazaki Y, Takahashi K: Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiol Genomics* 2003, 12(3):251-259.

21. Liang S, Li Y, Be X, Howes S, Liu W: Detecting and profiling tissue-selective genes. *Physiol Genomics* 2006, 26(2):158-162.

22. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG: Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput Biol* 2009, 5(6):e1000417.

23. Xu J, He Y, Qiang B, Yuan J, Peng X, Pan XM: A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics* 2008, 9:8.

24. Xue Y, Zhou F, Fu C, Xu Y, Yao X: SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* 2006, 34(Web Server issue):W254-257.

25. Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y: Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics* 2009, 9(12):3409-3412.

26. Huang LT, Gromiha MM, Ho SY: iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* 2007, 23(10):1292-1293.

27. Godzik A, Jambon M, Friedberg I: Computational protein function prediction: are we making progress? *Cell Mol Life Sci* 2007, 64(19-20):2505-2511.

28. Vakser IA, Kundrotas P: Predicting 3D structures of protein-protein complexes. *Curr Pharm Biotechnol* 2008, 9(2):57-66.

29. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P: Prediction of deleterious human alleles. *Hum Mol Genet* 2001, 10(6):591-597.

30. Sunyaev SR, Lathe WC, 3rd, Ramensky VE, Bork P: SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* 2000, 16(8):335-337.

31. Dimmic MW, Sunyaev S, Bustamante CD: Inferring SNP function using evolutionary, structural, and computational methods. *Pac Symp Biocomput* 2005:382-384.

32. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J: Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* 2003, 327(5):1021-1030.

33. Cheng TM, Lu YE, Vendruscolo M, Lio P, Blundell TL: Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol* 2008, 4(7):e1000135.

34. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999, 22(3):231-238.

35. Yue P, Moult J: Identification and analysis of deleterious human SNPs. *J Mol Biol* 2006, 356(5):1263-1274.

36. de Alencastro G, McCloskey DE, Kliemann SE, Maranduba CM, Pegg AE, Wang X, Bertola DR, Schwartz CE, Passos-Bueno MR, Sertie AL: New SMS mutation leads to a striking reduction in spermine synthase protein function and a severe form of Snyder-Robinson X-linked recessive mental retardation syndrome. *J Med Genet* 2008, 45(8):539-543.

37. Zhang Z, Teng S, Wang L, Schwartz CE, Alexov E: Computational analysis of missense mutations causing Snyder-Robinson syndrome. *Hum Mutat* 2010, 31(9):1043-1049.

CHAPTER TWO

A MACHINE LEARNING APPROACH FOR PREDICTING HUMAN TISSUE-SPECIFIC GENES USING MICROARRAY EXPRESSION DATA[1]

ABSTRACT

Understanding how genes are expressed specifically in particular tissues is a fundamental question in developmental biology. Many tissue-specific genes are involved in the pathogenesis of complex human diseases. However, experimental identification of tissue-specific genes is time consuming and difficult. The accurate predictions of tissue-specific gene targets could provide useful information for biomarker development and drug target identification. In this study, we have developed a machine learning approach for predicting the human tissue-specific/selective genes using microarray expression data. The lists of known tissue-specific genes for different tissues were collected from UniProt database, and the expression data retrieved from the previously compiled dataset according to the lists were used for input vector encoding. Random Forests (RFs) and Support Vector Machines (SVMs) were used to construct accurate classifiers. The RF classifiers were found to outperform SVM models for tissue-specific gene prediction. The results suggest that the candidate genes for brain or liver specific expression can provide valuable information for further experimental studies. Our approach was also applied for identifying tissue-selective gene targets for different types of tissues. The approach provides an efficient way to select some interesting genes for developing new biomedical markers and improve our knowledge of tissue-specific expression.

[1]Teng S, Wang L: A machine learning approach for predicting human tissue-specific genes using microarray expression data, in preparation.

BACKGROUND

Understanding how different tissues achieve specificity is a fundamental question in tissue ontogenesis and evolution. Some genes are highly expressed in a particular tissue and lowly expressed or not expressed in other tissues. These genes are generally called tissue-selective genes. The genes are responsible for specialized functions in particular tissues, and thus can serve as the biomarkers for specific biological processes. In addition, many tissue-selective genes are involved in the pathogenesis of complex human diseases [1], including insulin signaling pathways in diabetes [2] and tumour–host interactions in cancer [3]. Since the majority of disease genes have the tendency to be expressed preferentially in particular tissues [4], identifying tissue-selective genes is also important for drug target selection in biomedical research. Tissue-specific genes, which are expressed specifically in a particular tissue, are regarded as the special case of tissue selective genes. The identification of tissue-specific genes could help biologists to elucidate the molecular mechanisms of tissue development and provide valuable information for identifying candidate biomarkers and drug targets.

Different methods have been proposed to identify and characterize tissue-specific genes. Traditional experimental technologies, including RT-PCR and Northern blot, are usually carried out at the single-gene level and thus time-consuming. High-throughput technologies, such as Expressed Sequence Tag (EST) sequencing and DNA microarrays, have the capacity to perform genome-wide analysis with high efficiency. The DNA microarray technology can generate large amounts of gene expression data from various tissues, and provide the useful data source for analyzing tissue-specific genes. Several

statistical methods have been applied for identifying tissue-specific genes using gene expression data. Kadota and co-workers [5] described an unsupervised method to select the tissue-specific genes using Akaike's information criterion (AIC) approach. Another method called ROKU [6] has been developed by the same group for detecting tissue-specific gene expression patterns. The approach used Shannon entropy and outlier detection to scan expression profiles for ranking tissue-specific genes. Liang et al. [7] developed a statistical method based on hypothesis testing procedures to profile and identify the tissue-selective genes. However, the statistical methods for tissue-specific gene prediction suffer from drawbacks. The microarray expression data are generated from different experiments, both biological variations and experimental noise result in significant variations in data quality. The statistical methods usually assigned an equal weight to each observation for prediction. Thus, the methods do not work well for non-linear models and may not detect the hidden expression patterns from the noisy microarray data. Moreover, the statistical methods do not use biological knowledge for prediction. The simple data-driven analysis may produce some misleading results for further experimental studies.

Machine learning can automatically recognize hidden patterns in complex data. It has been shown that machine learning can be used to construct accurate classifiers for tissue-specific gene prediction. Chikina et al [8] used Support Vector Machines (SVMs) to predict tissue-specific gene expression in *Caenorhabditis elegans* with whole-animal microarray data. The SVM classifiers reached high predictive performances in nearly all tissues. It was shown that the approach outperformed clustering methods and provided

valuable information for further experimental studies. However, it is still unknown whether machine learning methods can be used to predict tissue-specific genes in human.

In a previous study [9], a large dataset has been compiled from a compendium of microarray expression profiles collected from 131 microarray datasets in different studies. The integrated dataset contained 2,968 expression profiles for various human tissues including brain (616 profiles) liver (117 profiles), testis (36 profiles), blood (409 profiles) and kidney (73 profiles). A computational method was developed for predicting tissue-selective genes with the integrated dataset using both microarray intensity values and detection calls. However, the method assigned an equal weight to each expression profile for prediction. In this study, a machine learning approach was developed for human tissue-specific gene prediction using the available dataset. According to the lists of known tissue-specific genes, the gene expression data were extracted from the compiled dataset and used for classifier construction. Random Forests (RFs) and Support Vector Machines (SVMs) were trained with the expression data to construct accurate classifiers. The results indicate that the RF classifiers achieved better predictive performance for tissue-specific gene prediction. The approach generated large numbers of candidate genes for brain and liver-specific expression. The examinations of high scoring genes suggest that our approach can be used to select candidate genes for experimental studies.

METHODS

A schematic diagram of the approach used in this study is shown in Figure 2.1. The microarray expression profiles of various human tissues were collected from NCBI GEO database [9]. The selected profiles were integrated into a single dataset through normalization and transformation. The lists of known tissue-specific genes were manually collected from UniProt database. The tissue-specific gene expression data were extracted from the integrated single dataset and labelled as positive training instances. The remaining expression data were randomly divided into two subsets. The negative dataset contained tenfold number of data instances as the positive instances. Random Forests (RFs) and Support Vector Machines (SVMs) were trained with the training instances to construct classifiers. The tenfold cross-validation method was performed to evaluate the classifier performance. The models were then used to score the remaining data instances for prediction. The classifier construction and prediction were repeated ten times, and the candidate genes were prioritized according to their average classifier outputs from ten predictions.

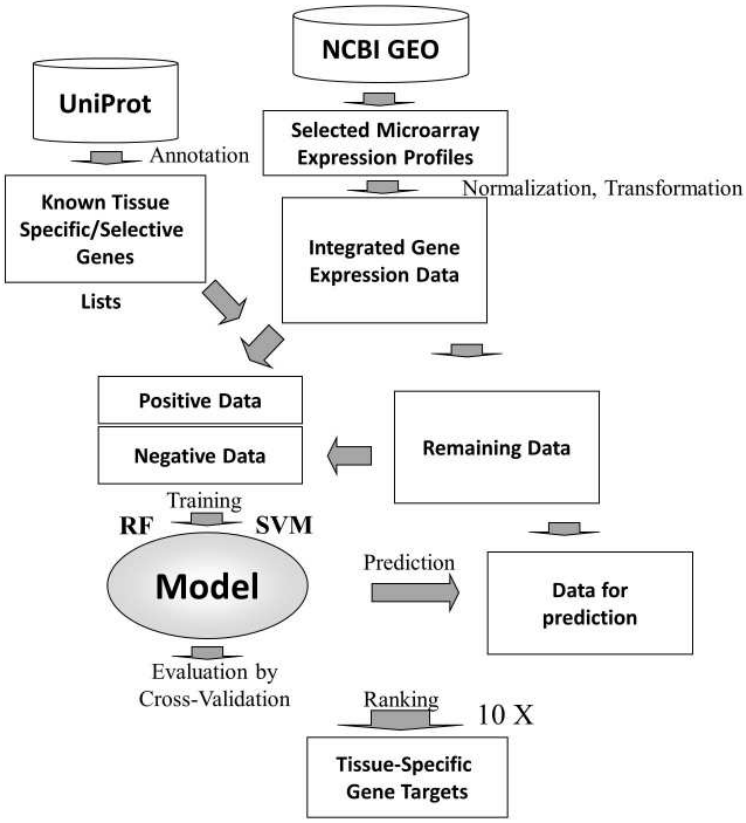

Figure 2.1 Schematic diagram of the approach for predicting tissue-specific genes.

Microarray data collection and integration

The approach for compiling human microarray expression profiles was described previously [9]. The microarray gene expression profiles of various human tissues from 131 microarray studies were compiled from NCBI GEO database. The expression profiles were generated using the Affymetrix HG-U133 Plus 2.0 Array and obtained from normal tissue samples. The raw data in CEL file format were organized into different normalization groups and normalized using the dChip tool with the invariant set method [10]. The outlier array exclusion and global median transformation were used to integrate

the normalized microarray profiles into a single dataset. The dataset used in this study contained 54,613 probe sets and 2,968 expression profiles.

Training data preparation

Tissue-selective genes are defined as the genes whose expression is enriched for one or a few similar tissue types. The genes were manually collected from the UniProt database. The particular tissue name was used as a query and the reviewed human genes were selected for preparation. The tissue-selective genes are defined as the genes that are expressed preferentially in a particular tissue from the descriptions of their annotations. Most of the genes are identified by the experimental methods, which are independent from the microarray expression data in the list. In this study, 408 brain-selective genes, 96 liver-selective genes, 326 testis-selective genes, 324 blood-selective genes and 45 kidney-selective genes were collected from UniProt database. Tissue-specific genes, whose expression is specific to only one particular tissue type, are considered as the special case of tissue-selective genes. 289 brain-specific genes and 69 liver-specific genes were selected from the corresponding tissue-selective genes with the annotation that their expression is specific to only brain or liver. Tissue specific-genes are the focus of the present study.

According to the known tissue-specific/selective gene lists, the tissue-specific/selective gene expression data was retrieved from the integrated microarray dataset and labelled as the positive training instances. The probe sets with detectable expression signals in corresponding tissue samples were selected for classifier

construction. For tissue-specific gene prediction, the expression values for 403 probe sets of brain-specific genes and 90 probe sets of liver-specific genes were used for input vector encoding. 692 probe sets of brain-selective genes, 150 probe sets of liver-selective, 430 probe sets of testis-selective genes, 456 probe sets of blood-selective genes and 76 probe sets of kidney-selective genes were used for tissue-selective gene prediction.

The negative examples were defined as the genes that do not have preferential expression in particular tissues. For this study, we randomly selected the data instances from the remaining data and labelled as the negative training instances. The number of negative instances was set as tenfold with positive instances to make enough data instances for training. The negative and positive data instances were combined as the training dataset to construct classifiers using machine learning algorithms. The remaining probes were used as the candidate genes for prediction with the classifiers constructed from the training dataset.

Random Forests

The use of 2,968 expression profiles for input vector encoding gives the same number of input variables. One potential problem is model overfitting since there were only a small number of positive instances (probe sets of known tissue-specific genes) available for this study. The problem could be solved using the Random Forests (RFs) learning algorithm. A typical RF model is made up many independent decision trees constructed using bootstrap samples from the training data. During tree construction, m variables out of all the n input variables ($m << n$) are randomly selected at each node, and

the tree node are split using the selected *m* variables. The RF classifier then combines the votes made by the decision trees, and gives the most popular class as the output of the ensemble for classification. Because of the random feature selection, RFs could handle a large number of input variables and avoid model overfitting. In the present study, the randomForest package in R was used for classifier construction. The number of variables selected to split each node (*mtry*) was set to 6, and the number of trees to grow (*ntree*) was set to 1000. Other values of the *mtry* and *ntree* parameters for training were also examined, but did not result in significant improvement of classifier performance.

Support vector machine training

Support Vector Machines (SVMs) are computational algorithms that can learn from training examples for binary classification. SVM classifiers were constructed and compared with RF classifiers for identifying human tissue-specific genes. The SVM learning algorithm can be described by four basic concepts, including the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function [11]. For a linear classification, the data instance in classifier is represented as an *n*-dimensional vector, and an $(n - 1)$ dimensional hyperplane is used to separate the positive instances from the negative ones. The SVMlight software package (http://svmlight.joachims.org/) was utilized to construct the SVM classifiers using the linear function in this study. The polynomial and radial basis function (RBF) kernel functions were also examined for classifier constructions, but the classifiers did not achieve high predictive performances in cross-validation tests.

Classifier evaluation and prediction

This study used a tenfold cross-validation method to evaluate classifier performance. Positive and negative instances were randomly distributed into ten folds. In each of the ten iterations, nine of the ten folds were used to train a classifier, and then the remaining one fold was used as the test data to evaluate the classifier. Since the dataset was imbalanced, the positive instances of training data were replicated to get the approximately equal number with the negative instances. However, the positive instances in the test data were not replicated. The predictions made for the test instances in all the ten iterations were combined and used to compute the following performance measures:

$$\text{Accuracy (AC)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$\text{Sensitivit y (SN)} = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{Specificiy (SP)} = \frac{TN}{TN + FP} \tag{2.3}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4}$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives. For imbalanced datasets, above measures could be misleading. Thus, the Receiver Operating Characteristic (ROC) curve [12], which generated by varying the output threshold of a classifier and plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) for each threshold value, was used for classifier evaluation and comparison. Since the ROC curve of an accurate classifier is close to the left-hand and top borders of

the plot, the area under the curve (AUC) can be used as a reliable measure of classifier performance [13]. The good classifiers have AUC values close to 1, whereas weak classifiers have AUC values near to 0.5.

The classifier construction and prediction were repeated ten times. In each run, the performance of classifier was computed by above measures. The mean value and standard deviation of the measures in ten runs were calculated to check the average levels and variations of classifier performances, respectively. The classifier was used to evaluate the candidate genes for prediction. The tissue-specific gene targets were sorted according to the decreasing average value of classifier outputs from ten predictions, and a higher value indicates a higher probability of being expressed predominantly in a particular tissue.

RESULTS AND DISCUSSION

Dataset validation

The known tissue-specific genes are expressed predominantly in particular tissues, so the transcripts of the genes were expected to be detected in corresponding tissue samples in the integrated microarray dataset. To visualize the expression patterns of the known tissue-specific genes, TM4 MeV [14] was used to generate the heat maps for brain and liver-specific genes. As shown in Figure 2.2, the known brain-specific genes have expression patterns in brain as well as retina samples. Since retina shares the common embryonic origins with brain and translates the visual images into nerve signals

for brain, the retina is considered as the sensory part of the brain. Thus, the known brain-specific genes may also have some expression levels in retina samples.

The transcripts of known liver-specific genes are detected clearly in liver samples (Figure 2.2). The results suggest that the expression data according to our lists of known tissue-specific genes can provide useful information for classifier construction using machine learning methods. It is noteworthy that some probe sets of known tissue-specific genes have high expression or no expression for all tissue samples. To improve the quality of classifiers, the probes without detectable expression signals in all the samples are excluded from the training dataset.

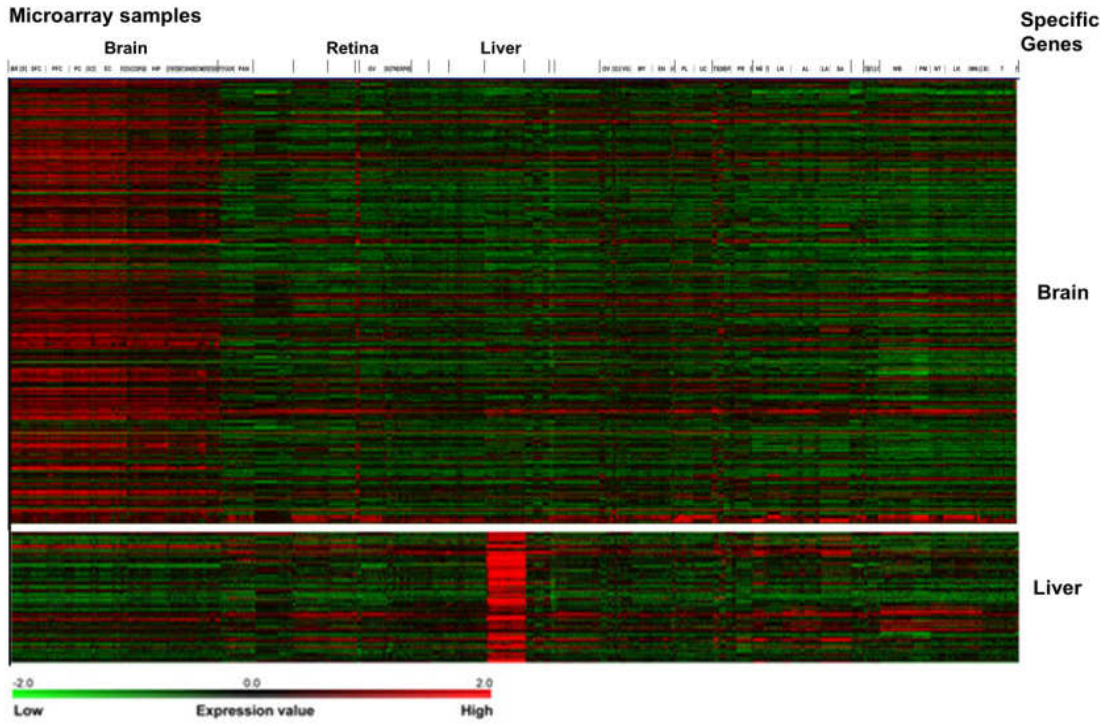

Figure 2.2 Visualization of known tissue-specific gene expression patterns.

Prediction of tissue-specific genes

Random Forests (RFs) and Support Vector Machines (SVMs) were used to construct classifiers for predicting brain and liver-tissue specific genes. The results suggest that RF classifiers reached better predictive performance than SVM models (Table 2.1 and Figure 2.3). We identified 1,408 brain-specific gene targets and 493 liver-specific gene targets using RF classifiers (Appendix B), which are even more than tissue-selective genes identified in the previous study (222 brain-selective targets and 69 liver-selective targets) [9]. It was shown that the transcripts of candidate genes could be detected clearly in corresponding tissue samples (Figure 2.4), and the functions of predicted targets were consistent with tissue origins in GO enrichment analysis (Table 2.2 and 2.4). High scoring gene targets with brain or liver-specific expression have been examined (Table 2.3 and 2.5), and the results suggest that the approach can provide useful information for identification of novel gene targets in biomedical research.

In this study, we constructed both RF and SVM classifiers for predicting brain and liver-specific genes. 403 probe sets of brain-specific genes and 90 probe sets of liver-specific genes were used for classifier construction. For brain-specific gene prediction, the RF classifier achieved the AUC value at 0.9488 (Table 2.1), which is significantly higher than the AUC value of SVM classifier (AUC = 0.8937). The RF classifier reached 53.73% sensitivity and 97.43% specificity, and MCC = 0.5676. For liver-specific gene prediction, the SVM classifier gave MCC = 0.8350 and ROC AUC = 0.9854. The RF classifier achieved a similar level of performance with MCC = 0.8290 and ROC AUC =

0.9777. Thus, the results suggest that the RF algorithm performs better for predicting tissue-specific genes in this study.

Table 2.1 Comparison of Random Forest and Support Vector Machine classifiers for predicting tissue-specific genes. The values outside and inside brackets are the average value and standard deviation of measures in ten classifier evaluations, respectively.

Tissue	Method	AC (%)	SN (%)	SP (%)	MCC	ROC AUC
Brain	**SVM**	92.07 (±0.302)	54.23 (±1.227)	95.82 (±0.263)	0.5091 (±0.015)	0.8937 (±0.003)
	RF	93.48 (±0.240)	53.73 (±1.485)	97.43 (±0.153)	0.5676 (±0.016)	0.9488 (±0.002)
Liver	**SVM**	97.29 (±0.421)	84.11 (±2.281)	98.61 (±0.309)	0.8350 (±0.025)	0.9854 (±0.004)
	RF	97.29 (±0.341)	79.00 (±1.355)	99.12 (±0.255)	0.8290 (±0.0213)	0.9777 (±0.002)

The ROC curves of RF and SVM classifiers for predicting brain-specific genes and live-specific genes have been compared in Figure 2.2. The ROC curves of RF and SVM classifiers are not significantly different for the prediction of liver-specific genes (Figure 2.2b). However, The ROC curve of RF classifier was clearly better than the SVM classifier for the prediction of brain-specific genes (Figure 2.2a). The results confirm that RF classifier outperforms the SVM models for tissue-specific gene prediction. The possible reason is that RFs can handle a large number of input variables and avoid model overfitting. The use of 2,968 expression profiles for input vector encoding results in the same large number of input variables, which may lead to model overfitting. Interestingly, the RF algorithm can handle the situation and show better predictive performance in the present study.

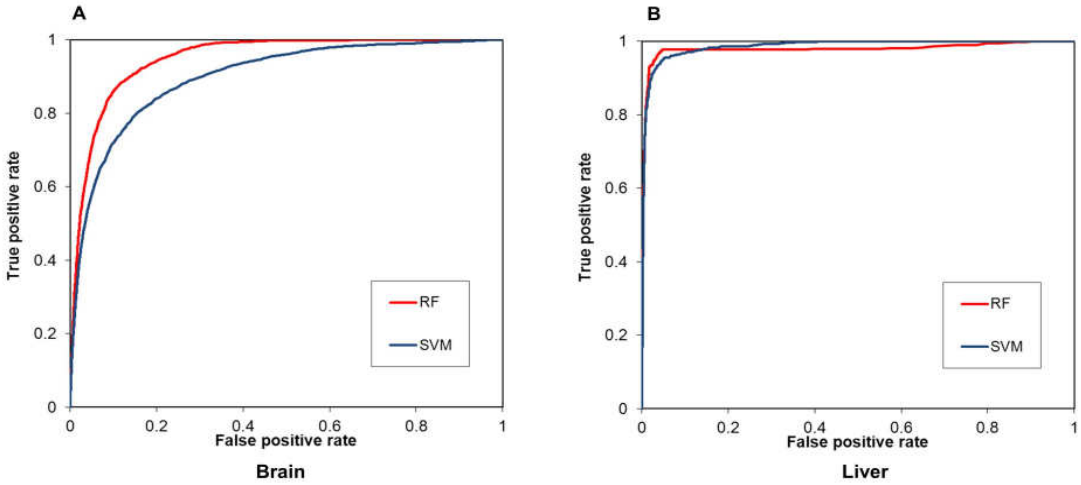

Figure 2.3 ROC curves to compare the performances of Random Forest (RF) and Support Vector Machine (SVM) classifiers for predicting tissue-specific genes.

Brain-specific gene expression

The human brain gives us the ability to think and sets us apart from other animals. It has a highly complex structure which contains different regions with specific functions. For example, the hippocampus is involved in spatial navigation and long-term memories, whereas the cerebral cortex plays key roles in language, attention and consciousness. Any damage in these regions results in various mental disorders including Alzheimer disease, Parkinson's disease and Mood disorder. In this study, the predicted brain-specific genes are expected to have preferential expression in the brain, and may play important roles in neuron functions such as synaptic transmission and neuronal migration.

In the study, 1,408 candidate targets with positive scores (the average value of classifier outputs from ten predictions) were predicted as the brain-specific genes (Additional file B1). In Figure 2.4, the expression patterns of candidate gene targets using RF classifier are visualized with the heat maps generated using TM4 MeV. The predicted

targets show clear expression patterns in brain samples, which indicates that our approach is useful for brain-specific gene prediction. Similar to the known brain-specific genes, the transcripts of the predicted targets are also detected in retina samples. GO enrichment analysis of the candidate targets demonstrates that many candidate genes have basic neuron functions (Table 2.2). For example, neurotransmission is an electrical or chemical signal motion within synapses caused by transmission of a nerve impulse. The predictions are enriched for neurotransmission-related GO terms such as “synapse”, “synapse part”, “transmission of nerve impulse”, “neuron projection”, “synaptic transmission” and “passive transmembrane transporter activity”. Some channel-related GO terms including “ion channel activity”, “substrate specific channel activity”, “gated channel activity” and “channel activity” are detected in the enrichment analysis of our predictions.

Table 2.2 GO term enrichment analysis of predicted brain-specific genes.

Category	**Term**	**Count***	**%***	**P-Value***
GOTERM_CC_FAT	GO:0045202~synapse	103	11.41	1.37E-49
GOTERM_CC_FAT	GO:0044456~synapse part	83	9.19	2.68E-45
GOTERM_BP_FAT	GO:0019226~transmission of nerve impulse	80	8.86	5.25E-36
GOTERM_CC_FAT	GO:0043005~neuron projection	85	9.41	4.00E-35
GOTERM_BP_FAT	GO:0007268~synaptic transmission	73	8.08	6.82E-35
GOTERM_MF_FAT	GO:0005216~ion channel activity	76	8.42	2.29E-30
GOTERM_MF_FAT	GO:0022838~substrate specific channel activity	77	8.53	3.03E-30
GOTERM_MF_FAT	GO:0022836~gated channel activity	68	7.53	4.80E-30
GOTERM_MF_FAT	GO:0015267~channel activity	77	8.53	3.28E-29
GOTERM_MF_FAT	GO:0022803~passive transmembrane transporter activity	77	8.53	3.87E-29

*Count: the number of genes involved in the given GO term; %: the percentage of involved genes in total genes; P-Value: the modified Fisher Exact P-Value.

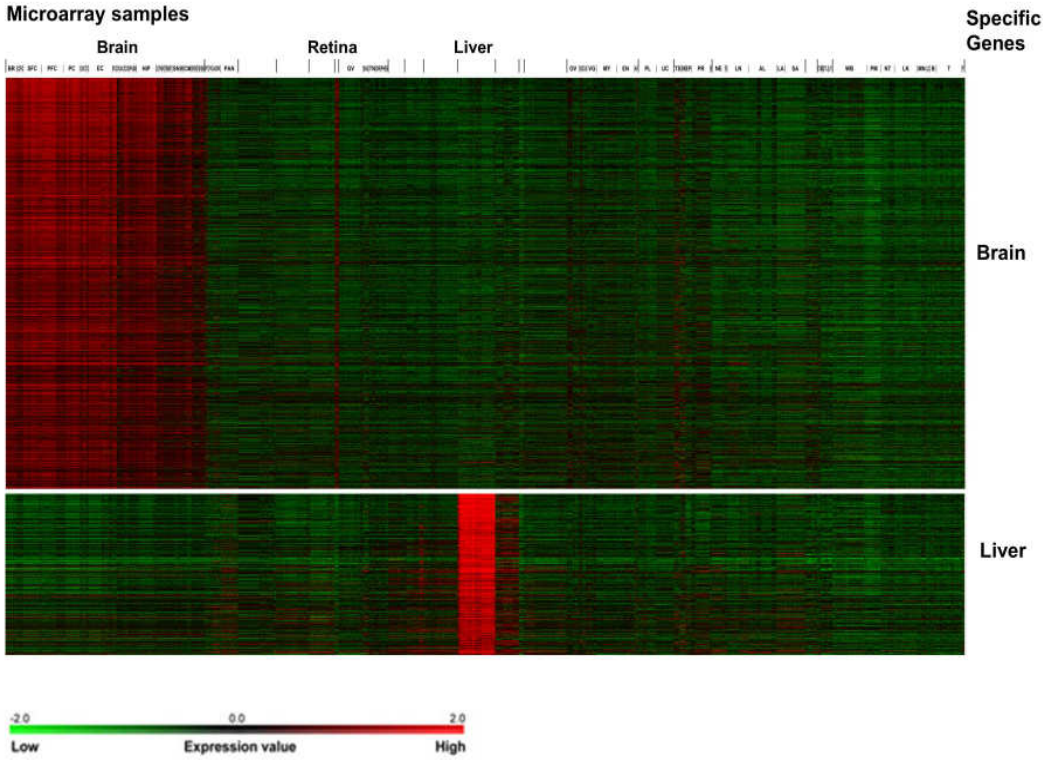

Figure 2.4 Visualization of predicted tissue-specific gene expression patterns.

Table 2.3 shows the top 20 high-scoring predictions from RF classifiers. The predicted targets were not shown to have brain-specific expression from the annotations of UniProt, thus the genes are excluded from the training datasets. However, recent studies suggest that some predicted targets, including *BRUNOL4*, *ANKS1B*, *TRIM9*, *NCAN*, *FAIM2*, *OPCML* and *FXYD7*, are expressed predominantly in the brain. For example, the RNA-binding protein encoded by *BRUNOL4* plays an important role in many cellular processes including RNA stability, pre-mRNA alternative splicing, mRNA editing and translation [15, 16]. It was shown that the protein was predominantly expressed in the brain with enrichment in the hippocampus [17]. In this study, the probes

of *BRUNOL4* have the highest (223654_s_at, 0.8753) and fourth-ranked (238966_at, 0.8600) scores. *ANKS1B* encodes an Amyloid-beta protein which can regulate the nucleoplasmic coilin protein interactions in neuronal cells. Previous studies showed that the protein was mainly expressed in brain and may be implicated in Alzheimer's disease [18]. Brain-specific E3 ligase encoded by *TRIM9* has a high level of expression in the cerebral cortex and may be involved in the pathogenesis of Parkinson's disease [19]. Neurocan (*NCAN*) modulates neuronal adhesion and migration and is expressed preferentially in the brain [20]. The protein encoded by *FAIM2* could protect cells from Fas-mediated apoptosis and shows a high level of expression in the hippocampus [21]. It was shown that *OPCML* was predominantly expressed in cerebellum and cerebral cortex [22], whereas *FXYD7* was preferentially expressed in the brain [23].

Other predicted targets have not been previously suggested to have brain-specific expression, but some candidate genes, including *GNAO1*, *SV2A*, *SYN2*, *UNC13A* and *NTRK2*, are involved in basic neuron functions (Table 2.3). Guanine nucleotide binding protein (*GNAO1*) mediates the physiological effects of various neuronal receptors [24]. *SV2A*, *SYN2* and *UNC13A* encode proteins which are important for synaptic transmission in the central and peripheral nervous system [25, 26]. *NTRK2* encodes a neurotrophic tyrosine kinase receptor for brain-derived neurotrophic factor (*BDNF*) and is implicated in childhood mood disorder [27]. By contrast, the functions of some high scoring genes in brain remain to be characterized. *HS6ST3* encodes a Heparan sulphate sulfotransferase which plays a key role in the modulation of fibroblast growth factor signalling [28]. The protein encoded by *SCN2A* forms a voltage-dependent sodium channel and is associated

with generalized epilepsy with febrile seizures plus [29]. The corresponding genes of three cDNA sequences (*LOC389073*, *AA879409*, *AI186173*) were not determined, and their functions in the brain are not clear. The results suggest that the machine learning approach developed in the present study can be used to identify some interesting targets for further experimental studies.

Table 2.3 List of high-scoring genes with specific expression in the brain.

Probe	**Gene**	**Description**	**Score***
223654_s_at	BRUNOL4	Bruno-like 4, RNA binding protein (Drosophila)	0.8753
227440_at	ANKS1B	Ankyrin repeat and sterile alpha motif domain containing 1B	0.8685
230280_at	TRIM9	Tripartite motif-containing 9	0.866
238966_at	BRUNOL4	Bruno-like 4, RNA binding protein (Drosophila)	0.8345
205143_at	NCAN	Neurocan	0.832
204762_s_at	GNAO1	Guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O	0.8201
232276_at	HS6ST3	Heparan sulfate 6-O-sulfotransferase 3	0.8186
203619_s_at	FAIM2	Fas apoptotic inhibitory molecule 2	0.8124
241998_at	LOC389073	Similar to RIKEN cDNA D630023F18	0.8074
206381_at	SCN2A	Sodium channel, voltage-gated, type II, alpha subunit	0.8021
203069_at	SV2A	Synaptic vesicle glycoprotein 2A	0.7998
1557256_a_at	AA879409	CDNA FLJ37672 fis, clone BRHIP2012059	0.797
229039_at	SYN2	Synapsin II	0.7956
242651_at	AI186173	Transcribed locus	0.7951
227453_at	UNC13A	unc-13 homolog A (C. elegans)	0.7888
203618_at	FAIM2	Fas apoptotic inhibitory molecule 2	0.7744
229463_at	NTRK2	Neurotrophic tyrosine kinase, receptor, type 2	0.7728
214111_at	OPCML	Opioid binding protein/cell adhesion molecule-like	0.7722
214376_at	AI263044	Clone 24626 mRNA sequence	0.7668
220131_at	FXYD7	FXYD domain containing ion transport regulator 7	0.7662

* Score: the average value of RF classifier outputs from ten predictions.

Liver-specific gene expression

The liver is a vital organ for human metabolism, and plays key roles in detoxification, plasma protein synthesis, glycogen storage and hormone production. For example, liver is the source and target organ of inflammatory mediators in the pathogenesis of inflammatory response syndrome [30], and it is responsible for the production of coagulation factors. Thus, the liver-specific targets identified in this study might be involved in basic liver functions. We identified 493 liver-specific gene targets with positive scores in the analysis (Additional file B2). The functional analysis of the liver-specific gene targets using RF classifier confirms that many of the predicted targets are enriched for liver-related GO terms (Table 2.4). For example, the GO terms for inflammatory response contained “acute inflammatory response”, “response to wounding” and “activation of plasma proteins involved in acute inflammatory response”; the coagulation-related GO terms included “blood coagulation”, “coagulation” and “hemostasis”. The expression patterns of the predicted liver-specific genes are visualized with the heat map (Figure 2.3). Clearly, the transcripts of the predicted targets are predominantly detected in liver samples.

Table 2.4 GO term enrichment analysis of predicted liver-specific genes.

Category	Term	Count*	%*	P-Value*
GOTERM_BP_FAT	GO:0002526~acute inflammatory response	29	8.41	1.65E-24
GOTERM_BP_FAT	GO:0009611~response to wounding	55	15.94	8.55E-23
GOTERM_CC_FAT	GO:0005615~extracellular space	63	18.26	8.65E-23
GOTERM_CC_FAT	GO:0005576~extracellular region	109	31.59	1.35E-21
GOTERM_BP_FAT	GO:0007596~blood coagulation	25	7.25	5.55E-19
GOTERM_BP_FAT	GO:0050817~coagulation	25	7.25	5.55E-19
GOTERM_BP_FAT	GO:0007599~hemostasis	25	7.25	2.33E-18
GOTERM_BP_FAT	GO:0055114~oxidation reduction	54	15.65	2.46E-18
GOTERM_BP_FAT	GO:0006956~complement activation	18	5.22	2.70E-18
GOTERM_BP_FAT	GO:0002541~activation of plasma proteins involved in acute inflammatory response	18	5.22	4.37E-18

*Count: the number of genes involved in the given GO term; %: the percentage of involved genes in total genes; P-Value: the modified Fisher Exact P-Value.

As listed in Table 2.5, 17 of the top 20 high-scoring genes are involved in the metabolism of human liver. The genes include *F11*, *F9*, *SERPINC1*, *APOA2*, *AKR1D1*, *ACSM2*, *ITIH2*, *PON1*, *CPB2*, *AFM*, *NR0B2*, *ALB*, *CYP4A11*, *PGLYRP2* and *SLC22A7*. For example, *F11*, *F9* and *SERPINC1* are involved in the regulation of blood coagulation cascade [31]. *APOA2* encodes apolipoprotein which is synthesized mainly in liver and involved in the metabolism of high density lipoprotein [32]. *AKR1D1* encodes the aldo-keto reductase catalyzing the reduction of steroid hormones [33], whereas *ACSM2* encodes enzyme catalyzing the activation of medium-chain length fatty acids [34]. The genes were not shown to have liver-specific expression in UniProt annotations, but recent studies suggest that the genes are expressed preferentially in the liver. The expression and functions of other three predictions (*BG398937*, *C6* and *LPA*) have not been well documented in the literature.

Table 2.5 List of high-scoring genes with specific expression in the liver.

Probe	Gene	Description	Score*
206610_s_at	F11	Coagulation factor XI (plasma thromboplastin antecedent)	0.7869
1554491_a_at	SERPINC1	Serpin peptidase inhibitor, clade C member 1	0.7737
219465_at	APOA2	Apolipoprotein A-II	0.7609
217512_at	BG398937	Unknown	0.7559
207102_at	AKR1D1	Aldo-keto reductase family 1, member D1	0.7466
207218_at	F9	Coagulation factor IX	0.725
210168_at	C6	Complement component 6	0.7239
204987_at	ITIH2	Inter-alpha (globulin) inhibitor H2	0.7191
209978_s_at	LPA/PLG	Lipoprotein, Lp(a) / plasminogen	0.7191
214069_at	ACSM2	Acyl-CoA synthetase medium-chain family member 2	0.7099
206345_s_at	PON1	Paraoxonase 1	0.7004
206651_s_at	CPB2	Carboxypeptidase B2 (plasma)	0.6959
241914_s_at	ACSM2	Acyl-CoA synthetase medium-chain family member 2	0.6945
206840_at	AFM	Afamin	0.6846
206410_at	NR0B2	Nuclear receptor subfamily 0, group B, member 2	0.6837
214842_s_at	ALB	Albumin	0.6809
217319_x_at	CYP4A11	Cytochrome P450, family 4,subfamily A, polypeptide 11	0.6772
242817_at	PGLYRP2	Peptidoglycan recognition protein 2	0.6765
207407_x_at	CYP4A11	Cytochrome P450, family 4, subfamily A, polypeptide 11	0.6752
231398_at	SLC22A7	Solute carrier family 22, member 7	0.6746

* Score: the average value of RF classifier outputs from ten predictions.

Tissue-selective gene prediction

Tissue-specific genes are considered as the special case of tissue-selective genes. Our approach was developed for tissue-specific gene predictions, but its application to tissue-selective gene predictions is straightforward. In this study, the RF classifiers were used to predict the genes that are expressed preferentially in the brain, liver, testis, blood

and kidney. As shown in Table 2.6, The RF classifiers reached high predictive performance for tissue-selective gene prediction. For example, the classifier for brain-selective gene prediction shows overall accuracy (AC) at 92.70% with Matthews Correlation Coefficient (MCC) = 0.4925. The classifier for liver-selective gene prediction gave predictive performance with the overall accuracy at 96.02% and MCC = 0.7378. It is noteworthy that the classifiers used for tissue-specific gene prediction achieved higher predictive performance than those for tissue-selective gene prediction. For instance, the AUC value of RF classifier for brain-specific gene prediction (AUC = 0.9488, Table 2.1) is higher than that for brain-selective gene prediction (AUC = 0.9178, Table 2.6), whereas the RF classifier gave better predictive performance for liver-specific gene prediction (AUC = 0.9777, Table 2.1) than liver-selective gene prediction (AUC = 0.9547, Table 2.6). The possible explanation is that the tissue-specific genes are expressed specifically in only one particular tissue type, thus the clear expression patterns of the genes may improve the quality of classifiers and result in high predictive performance for predictions.

The RF classifiers gave high predictive performance for predicting genes that have preferential expression in other tissue types. The testis is the male sex gland, which produces sperm, male reproductive cell and sex hormones. The classifier for testis-selective gene prediction reached predictive performance with overall accuracy at 91.00% and ROC AUC = 0.8433. The blood transports oxygen and nutrients to other tissues and carries away waste products from cells. The classifier for blood-selective gene prediction showed overall accuracy at 93.29% with MCC = 0.5109 and ROC AUC =

0.9170. The kidneys play key roles in urinary system. The organs filter waste products from the blood and excrete them in urine. The classifier for kidney-selective gene prediction achieved predictive performance with overall accuracy at 93.62% with MCC = 0.4648 and ROC AUC = 0.9300. The results suggest that our approach can be used to identify the genes that have preferential expression in different types of tissues.

Table 2.6 Random Forest classifiers for predicting tissue-selective genes. The values outside and inside brackets are the average value and standard deviation of measures in ten classifier evaluations, respectively.

Tissue	**AC (%)**	**SN (%)**	**SP (%)**	**MCC**	**ROC AUC**
Brain	92.70 (±0.273)	43.55 (±1.212)	97.60 (±0.211)	0.4925 (±0.018)	0.9178 (±0.002)
Liver	96.02 (±0.341)	65.6 (±2.499)	99.07 (±0.191)	0.7378 (±0.024)	0.95467 (±0.003)
Testis	91.00 (±0.033)	1.49 (±0.405)	99.95 (±0.038)	0.0980 (±0.014)	0.8433 (±0.004)
Blood	93.29 (±0.190)	40.20 (±1.291)	98.53 (±0.108)	0.5109 (±0.016)	0.9170 (±0.002)
Kidney	93.62 (±0.508)	26.43 (±5.355)	99.73 (±0.159)	0.4648 (±0.062)	0.9300 (±0.003)

CONCLUSION

A machine learning approach has been developed in this study for identifying the human tissue/specific gene targets. Random Forests (RFs) and Support Vector Machines (SVMs) were trained separately with the microarray gene expression data to construct classifiers for prediction. It was shown that the RF classifiers outperform SVM models for tissue-specific gene prediction. 1,408 brain-specific gene targets and 493 liver-specific gene targets were identified using RF classifiers. The predicted targets show

clear expression patterns in corresponding tissue samples and have functions consistent with the tissues in GO enrichment analysis. The analysis of high-scoring candidate genes for brain and liver specific expression suggests that our approach can select some interesting targets for further experimental studies. Our approach could also provide useful information for tissue-selective gene prediction. The approach can be used to develop new drug targets for biomedical research and expand our knowledge of tissue-specific expression.

REFERENCES

1. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S: A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 2008, 105(52):20870-20875.

2. Saltiel AR, Kahn CR: Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* 2001, 414(6865):799-806.

3. Liotta LA, Kohn EC: The microenvironment of the tumour-host interface. *Nature* 2001, 411(6835):375-379.

4. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: The human disease network. *Proc Natl Acad Sci U S A* 2007, 104(21):8685-8690.

5. Kadota K, Nishimura S, Bono H, Nakamura S, Hayashizaki Y, Okazaki Y, Takahashi K: Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiol Genomics* 2003, 12(3):251-259.

6. Kadota K, Ye J, Nakai Y, Terada T, Shimizu K: ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinformatics* 2006, 7:294.

7. Liang S, Li Y, Be X, Howes S, Liu W: Detecting and profiling tissue-selective genes. *Physiol Genomics* 2006, 26(2):158-162.

8. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG: Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput Biol* 2009, 5(6):e1000417.

9. Wang L, Srivastava AK, Schwartz CE: Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics* 2010, 11 Suppl 2:S15.

10. Li C, Wong WH: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 2001, 98(1):31-36.

11. Noble WS: What is a support vector machine? *Nat Biotechnol* 2006, 24(12):1565-1567.

12. Swets JA: Measuring the accuracy of diagnostic systems. *Science* 1988, 240(4857):1285-1293.

13. Bradley A: The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997, 30:1145-1159.

14. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J: TM4 microarray software suite. *Methods Enzymol* 2006, 411:134-193.

15. Anant S, Henderson JO, Mukhopadhyay D, Navaratnam N, Kennedy S, Min J, Davidson NO: Novel role for RNA-binding protein CUGBP2 in mammalian RNA editing. CUGBP2 modulates C to U editing of apolipoprotein B mRNA by interacting with apobec-1 and ACF, the apobec-1 complementation factor. *J Biol Chem* 2001, 276(50):47338-47351.

16. Mukhopadhyay D, Houchen CW, Kennedy S, Dieckgraefe BK, Anant S: Coupled mRNA stabilization and translational silencing of cyclooxygenase-2 by a novel RNA binding protein, CUGBP2. *Mol Cell* 2003, 11(1):113-126.

17. Yang Y, Mahaffey CL, Berube N, Maddatu TP, Cox GA, Frankel WN: Complex seizure disorder caused by Brunol4 deficiency in mice. *PLoS Genet* 2007, 3(7):e124.

18. Ghersi E, Noviello C, D'Adamio L: Amyloid-beta protein precursor (AbetaPP) intracellular domain-associated protein-1 proteins bind to AbetaPP and modulate its processing in an isoform-specific manner. *J Biol Chem* 2004, 279(47):49105-49112.

19. Tanji K, Kamitani T, Mori F, Kakita A, Takahashi H, Wakabayashi K: TRIM9, a novel brain-specific E3 ubiquitin ligase, is repressed in the brain of Parkinson's disease and dementia with Lewy bodies. *Neurobiol Dis* 2010, 38(2):210-218.

20. Rauch U, Karthikeyan L, Maurel P, Margolis RU, Margolis RK: Cloning and primary structure of neurocan, a developmentally regulated, aggregating chondroitin sulfate proteoglycan of brain. *J Biol Chem* 1992, 267(27):19536-19547.

21. Somia NV, Schmitt MJ, Vetter DE, Van Antwerp D, Heinemann SF, Verma IM: LFG: an anti-apoptotic gene that provides protection from Fas-mediated cell death. *Proc Natl Acad Sci U S A* 1999, 96(22):12667-12672.

22. Reed JE, Dunn JR, du Plessis DG, Shaw EJ, Reeves P, Gee AL, Warnke PC, Sellar GC, Moss DJ, Walker C: Expression of cellular adhesion molecule 'OPCML' is down-regulated in gliomas and other brain tumours. *Neuropathol Appl Neurobiol* 2007, 33(1):77-85.

23. Tipsmark CK: Identification of FXYD protein genes in a teleost: tissue-specific expression and response to salinity change. *Am J Physiol Regul Integr Comp Physiol* 2008, 294(4):R1367-1378.

24. Kest B, Smith SB, Schorscher-Petcu A, Austin JS, Ritchie J, Klein G, Rossi GC, Fortin A, Mogil JS: Gnao1 (G alphaO protein) is a likely genetic contributor to variation in physical dependence on opioids in mice. *Neuroscience* 2009, 162(4):1255-1264.

25. Li L, Chin LS, Greengard P, Copeland NG, Gilbert DJ, Jenkins NA: Localization of the synapsin II (SYN2) gene to human chromosome 3 and mouse chromosome 6. *Genomics* 1995, 28(2):365-366.

26. Portela-Gomes GM, Lukinius A, Grimelius L: Synaptic vesicle protein 2, A new neuroendocrine cell marker. *Am J Pathol* 2000, 157(4):1299-1309.

27. Adams JH, Wigg KG, King N, Burcescu I, Vetro A, Kiss E, Baji I, George CJ, Kennedy JL, Kovacs M, Barr CL: Association study of neurotrophic tyrosine kinase receptor type 2 (NTRK2) and childhood-onset mood disorders. *Am J Med Genet B Neuropsychiatr Genet* 2005, 132B(1):90-95.

28. Kamimura K, Fujise M, Villa F, Izumi S, Habuchi H, Kimata K, Nakato H: Drosophila heparan sulfate 6-O-sulfotransferase (dHS6ST) gene. Structure, expression, and function in the formation of the tracheal system. *J Biol Chem* 2001, 276(20):17014-17021.

29. Sugawara T, Tsurubuchi Y, Agarwala KL, Ito M, Fukuma G, Mazaki-Miyazaki E, Nagafuji H, Noda M, Imoto K, Wada K, Mitsudome A, Kaneko S, Montal M, Nagata K, Hirose S, Yamakawa K: A missense mutation of the Na+ channel alpha II subunit gene Na(v)1.2 in a patient with febrile and afebrile seizures causes channel dysfunction. *Proc Natl Acad Sci U S A* 2001, 98(11):6384-6389.

30. Szabo G, Romics L, Jr., Frendl G: Liver in sepsis and systemic inflammatory response syndrome. *Clin Liver Dis* 2002, 6(4):1045-1066, x.

31. Kalafatis M, Egan JO, van 't Veer C, Cawthern KM, Mann KG: The regulation of clotting factors. *Crit Rev Eukaryot Gene Expr* 1997, 7(3):241-280.

32. Zhang T, Yao S, Wang P, Yin C, Xiao C, Qian M, Liu D, Zheng L, Meng W, Zhu H, Liu J, Xu H, Mo X: Apoa-Ii Directs Morphogenetic Movements of Zebrafish Embryo by Preventing Chromosome Fusion During Nuclear Division in Yolk Syncytial Layer. *J Biol Chem* 2011.

33. Charbonneau A, The VL: Genomic organization of a human 5beta-reductase and its pseudogene and substrate selectivity of the expressed enzyme. *Biochim Biophys Acta* 2001, 1517(2):228-235.

34. Boomgaarden I, Vock C, Klapper M, Doring F: Comparative analyses of disease risk genes belonging to the acyl-CoA synthetase medium-chain (ACSM) family in human liver and cell lines. *Biochem Genet* 2009, 47(9-10):739-748.

CHAPTER THREE

PREDICTING PROTEIN SUMOYLATION SITES FROM SEQUENCE FEATURES[2]

ABSTRACT

Protein sumoylation is a post-translational modification that plays an important role in a wide range of cellular processes. Small ubiquitin-related modifier (SUMO) can be covalently and reversibly conjugated to the sumoylation sites of target proteins, many of which are implicated in various human genetic disorders. The accurate prediction of protein sumoylation sites may help biomedical researchers to design their experiments and understand the molecular mechanism of protein sumoylation. In this study, a new machine learning approach has been developed for predicting sumoylation sites from protein sequence information. Random Forests (RFs) and Support Vector Machines (SVMs) were trained with the data collected from the literature. Domain-specific knowledge in terms of relevant biological features was used for input vector encoding. It was shown that RF classifier performance was affected by the sequence context of sumoylation sites, and twenty residues with the core motif ΨKXE in the middle appeared to provide enough context information for sumoylation site prediction. The RF classifiers were also found to outperform SVM models for predicting protein sumoylation sites from sequence features. The results suggest that the machine learning approach gives rise to more accurate prediction of protein sumoylation sites than previous studies. The RF and SVM models were used to develop a new web server, called seeSUMO (freely available

[2]Teng S, Luo H, Wang L: Predicting protein sumoylation sites from sequence features, submitted.

at http://bioinfo.ggc.org/seesumo), for sequence-based prediction of protein sumoylation sites.

BACKGROUND

Post-translational modifications regulate protein functions, and orchestrate a variety of cellular processes. Protein sumoylation, an important reversible post-translational modification, is essential for many eukaryotic cellular processes, including DNA damage recovery regulation, subcellular transport, transcription factor transactivation, protein stability, cell cycle progression and chromosome segregation [1]. Small ubiquitin-related modifier (SUMO) can be covalently attached to and detached from specific lysine residues in target proteins [2]. Many sumoylated proteins, including huntingtin, DJ-1, ataxin-1 and tau, play key roles in disease states. For instance, the stability and correct targeting of huntingtin are controlled by sumoylation, and any alternations of the process may cause Huntington's disease [3]. Sumoylation is also involved in the pathogenesis of Parkinson's disease, Alzheimer's disease, neuronal intranuclear inclusion disease, amyotrophic lateral sclerosis, spinobulbar muscular atrophy, spinocerebellar ataxia type 1 and several human cancers [4].

Only one or a few lysine residues in a protein may be involved in sumoylation. It is rather difficult and time-consuming to identify the sumoylated lysine among many candidate lysine residues through experimental approaches. Accurate computational prediction of protein sumoylation sites can help biologists better design their experiments and interpret the experimental data. A core consensus motif ΨKXE has been identified

for sumoylation sites, in which Ψ represents an aliphatic amino acid (I, V, L, A, P or M), K is the sumoylation site, X indicates any amino acid, and E is glutamic acid. Extended sumoylation motifs have also been reported [5], such as NDSM (negatively charged amino acid-dependent sumoylation motif: ΨKXE + downstream cluster of [D/E]) [6], PDSM (phosphorylation-dependent sumoylation motif: ΨKXEXXSP) [7] and SUMO-acetyl switch (ΨKXEP) [8]. These findings suggest that the sequence flanking the core motif (ΨKXE) may also contribute to the specific recognition of the sumoylation sites. Moreover, it is noteworthy that some sumoylation sites do not follow the above motifs, and not all lysine residues matched to these motifs are sumoylated. It is still challenging to accurately predict the true sumoylation sites recognized by the cellular machinery.

Accurate prediction of sumoylation site could help understand the mechanism of protein sumoylation underlying human genetic disorders. Several computational methods have been reported for predicting sumoylation sites. A statistical method used by the SUMOpre web server [9] can predict sumoylation sites at the overall accuracy of 96.71% and Matthews Correlation Coefficient of 0.6364 in cross-validation tests. Xue et al. [10] developed the SUMOsp 1.0 web server, which used the Group-based Phosphorylation Scoring (GPS) algorithm with the pattern recognition strategy MotifX for sumoylation site prediction. SUMOsp 2.0 [11] was developed with the upgraded GPS algorithm. It has been shown that SUMOsp 2.0 reached better predictive performance than SUMOsp 1.0. However, the previous studies did not utilize the domain-specific knowledge for classifier construction.

Domain-specific knowledge in terms of relevant biological features can be used to enhance classifier performance for predicting DNA-binding residues and protein stability changes upon amino acid substitutions [12-14]. For example, the predictive performance of DNA-binding site prediction could be significantly improved by using biochemical features [14] and evolutionary information [15] for input vector encoding. In this study, we have developed a new approach for sequence-based prediction of protein sumoylation sites using Random Forests (RFs) and Support Vector Machines (SVMs). The biological knowledge in terms of forty sequence features were used for input encoding. It was found that the RF classifier performance was affected by sequence context of sumoylation sites. The results obtained in this study indicate that the RF classifiers achieved better predictive performance than the SVM classifiers and previous predictors. To make our classifiers publicly accessible to the biological research community, we have developed a new web server called seeSUMO (freely available at http://bioinfo.ggc.org/seesumo).

METHODS

Data

We collected 457 experimentally verified sumoylation sites in 263 proteins, by searching the research articles in NCBI PubMed using 'SUMO' and 'sumoylation' as keywords (Appendix C, Table C.1). This dataset contained all the instances used by SUMOpre [9], including 268 sumoylation sites in 159 proteins from research articles reported before August 10, 2006. The other 189 sumoylation sites have been manually collected from research articles published between August 10, 2006 and June 1, 2010.

The amino acid sequences of these proteins were extracted from the SwissProt database. In order to remove redundancy in the dataset, the blastclust program in the BLAST software package (http://blast.ncbi.nlm.nih.gov/) was used for clustering analysis with a 25% sequence identity threshold, and ClustalX [16] was used for multiple sequence alignment of the sequences in each cluster. The redundant sumoylation sites were manually removed from the dataset. The final dataset contains 9,952 lysine residues in 247 proteins, including 425 non-redundant sumoylation sites used as positive data instances and 9,527 non-sumoylated lysine sites used as negative data instances.

To compare the predictive performance of our classifiers with previous predictors, the final dataset was divided into two subsets. The training dataset included 377 sumoylation sites and 8,237 non-sumoylated lysine residues in 221 proteins from publications before January 2010. The remaining 48 sumoylation sites and 1,290 non-sumoylation sites in 26 proteins reported after January 2010 were used as the test dataset for classifier evaluation and comparison.

Sequence logos

Protein Sequence Logos (http://www.cbs.dtu.dk/~gorodkin/appl/plogo.html) was used to generate the sequence logo for visualizing the sequence pattern of sumoylation motifs. The twenty eight residues with the core motif ΨKXE in the middle of 388 known sumoylation sites was used as the inputs, and the frequencies of residues at each position were measured in bits of information as described in previous studies [17, 18]. The

height of residue k at position i (d_{ik}) is proportional to its frequency relative to the expected frequencies, which is computed as follows:

$$d_{ik} = \frac{q_{ik} / p_k}{\sum_l q_{il} / p_l} I_i \qquad (3.1)$$

where q_{ik} represents the fraction of residue k at position i, and p_k indicates the priori amino acid distribution, which was set to the amino acid composition obtained from UniProtKB/Swiss-Prot Release 57.15 in this study. I_i is the information content of position i as described below:

$$I_i = \sum_{k \in A} I_{ik} = \sum_{k \in A} q_{ik} \log_2 \frac{q_{ik}}{p_k} \qquad (3.2)$$

where A is the set of residues including gaps.

Sequence features

Forty biological features, including ten biochemical features, seven structural features, nine thermodynamic features, six empirical features and eight other biological features, selected from Protscale [19] and AAindex [20], were used to encode each amino acid residue in a data instance (Appendix C, Table C.2). These features represent different types of biological knowledge such as biochemical properties, structural information, protein stability, folding energy, etc. For example, the biochemical feature, polarity (P), represents the dipole-dipole intermolecular interactions between the positively and negatively charged residues, and the structural feature, conformational parameter for alpha-helix (A), indicates the tendency of an amino acid to form the secondary structures, alpha-helix. Some of these features were used for predicting DNA-

binding residues and protein stability changes upon amino acid substitutions in previous studies [12-14].

Evolutionary Information

It was shown that utilizing the evolutionary information in terms of position-specific scoring matrix (PSSM) scores could improve the performance of Random Forests for DNA-binding site prediction [21]. The PSSM scores generated by PSI-BLAST in this study indicated how well each position of a sequence was conserved among its homologues. The protein sequences downloaded from UniProtKB/Swiss-Prot (http://www.pir.uniprot.org/, release 57.15) were used as the reference database, and PSI-BLAST was run for three iterations with the E-value threshold set to 1e-5.

Random Forests

The use of forty biological features and evolutionary information for input vector encoding gives rise to a large number of input variables, especially with a large window size. Considering the relatively small number of positive instances (experimentally identified sumoylation sites) available for this study, this might result in model overfitting. To avoid model overfitting, the Random Forest (RF) learning algorithm was used in this study. A typical RF model is made up of many independent decision trees constructed using bootstrap samples from the training data. During tree construction, m variables out of all the n input variables ($m << n$) are randomly selected at each node, and the tree nodes are split using the selected m variables. For classifying a data instance, a

RF classifier combines the votes made by the decision trees, and gives the most popular class as the output of the ensemble. Because of the random feature selection, RFs have the capability of handling a large number of input variables and avoid model overfitting.

In this study, the RF algorithm is implemented using the randomForest package in R. The number of variables selected to split each node (*mtry*) was set to 6, and the number of trees to grow (*ntree*) was set to 1000. Other values of the *mtry* and *ntree* parameters for training were also examined, but did not result in significant improvement of classifier performance.

Support vector machine training

Support Vector Machine (SVM) classifiers were also constructed, and compared with RF classifiers for protein sumoylation site prediction. The SVM method is a data-driven approach for binary classification. The SVM learning algorithm can be described by four basic concepts, including the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function [22]. For a linear classification, data instances are represented as n-dimensional vectors, and an $(n - 1)$ dimensional hyperplane is used to separate the positive instances from the negative ones. Non-linear classifications are generally used for the analysis of complex biological data. In these cases, a kernel function can be used to measure the distance between data points in a higher dimensional space, which allows the SVM algorithm to fit the maximum-margin hyperplane in the transformed space. The SVMlight software package

(http://svmlight.joachims.org/) was utilized to construct the SVM classifiers using the radial basis function (RBF) kernel in this study.

In this study, forty biological features were used to build the SVM models. However, the features used for classifier construction might contain redundant or correlated information. Thus, feature selection was performed to choose an optional subset of relevant features for constructing simple, efficient models. The five relevant features were selected by Random Forests, and then used to construct SVM classifiers.

Classifier evaluation

The predictive performance of classifier was evaluated by tenfold cross-validation. The whole dataset were randomly distributed into ten folds. In each of the ten iterations, the classifier was trained in nine of the ten folds and tested in the remaining one fold. Since the dataset was imbalanced with only 4% of lysine residues as sumoylation sites, the positive instances of training data were replicated to get the approximately equal number with the negative instances. However, the positive instances in the test data were not replicated. The prediction results made for the test data instances in all the ten iterations were combined and evaluated by various performance measures, including Accuracy (AC), Sensitivity (SN), Specificity (SP), Strength (ST) and Matthews Correlation Coefficient (MCC):

$$\text{Accuracy (AC)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$\text{Sensitivity (SN)} = \frac{TP}{TP + FN} \tag{3.4}$$

$$\text{Specificiy (SP)} = \frac{TN}{TN + FP} \tag{3.5}$$

$$\text{Strength (ST)} = \frac{\text{Sensitivit y} + \text{Specificit y}}{2} \tag{3.6}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives. For imbalanced datasets, the accuracy alone could be misleading. Thus, sensitivity, specificity and their average (strength) were also computed from prediction results. MCC was used to measure the correlation between predictions and the actual class labels. However, different trade-offs of sensitivity and specificity may give rise to different MCC values for a classifier.

The Receiver Operating Characteristic (ROC) curve [23] is probably the most robust approach for classifier evaluation and comparison. In the present study, the ROC curve was generated by varying the output threshold of a RF classifier and plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) for each threshold value. Since the ROC curve of an accurate classifier is close to the left-hand and top borders of the plot, the area under the curve (AUC) can be used as a reliable measure of classifier performance [24]. The range of AUC value is 0.5 (random guessing) to 1 (perfect classifier).

RESULTS AND DISCUSSION

Sequence patterns of protein sumoylation sites

Protein sumoylation sites are often identified with the consensus motif ΨKXE, where Ψ represents an aliphatic amino acid (I, V, L, A, P or M), and X indicates any residue. However, 159 (~35%) of 457 known sumoylation sites in this study do not contain the core motif (ΨKXE), whereas the dataset contains 228 non-sumoylated lysine residues that match this motif. To visualize the sequence patterns in sumoylation sites and their flanking sequences, the sequence logo was generated using the 28-residue sequences from 388 experimentally identified sumoylation sites (Figure 3.1). The result suggests that certain positions outside of the core motif (ΨKXE), such as the positions -7, 1, 3, 4 and 9, may contain some information for the specific recognition of sumoylation sites in the cell. For instance, the most abundant residue at positions 1 is Proline (P), which agrees well with the SUMO-acetyl switch (ΨKXEP). Interestingly, glutamine (Q), Methionine (M) and Threonine (T) appear to be more abundant than any other residues in the X position of the core motif ΨKXE, suggesting that there may be subtle amino acid preference at the X position.

The above observations suggest that the flanking sequences of protein sumoylation sites have some subtle patterns. However, these patterns may not be modelled completely by consensus motifs or sequence logos, which do not consider the dependence among neighboring residues. Thus, a machine learning approach has been developed in this study to model the sequence patterns of protein sumoylation sites.

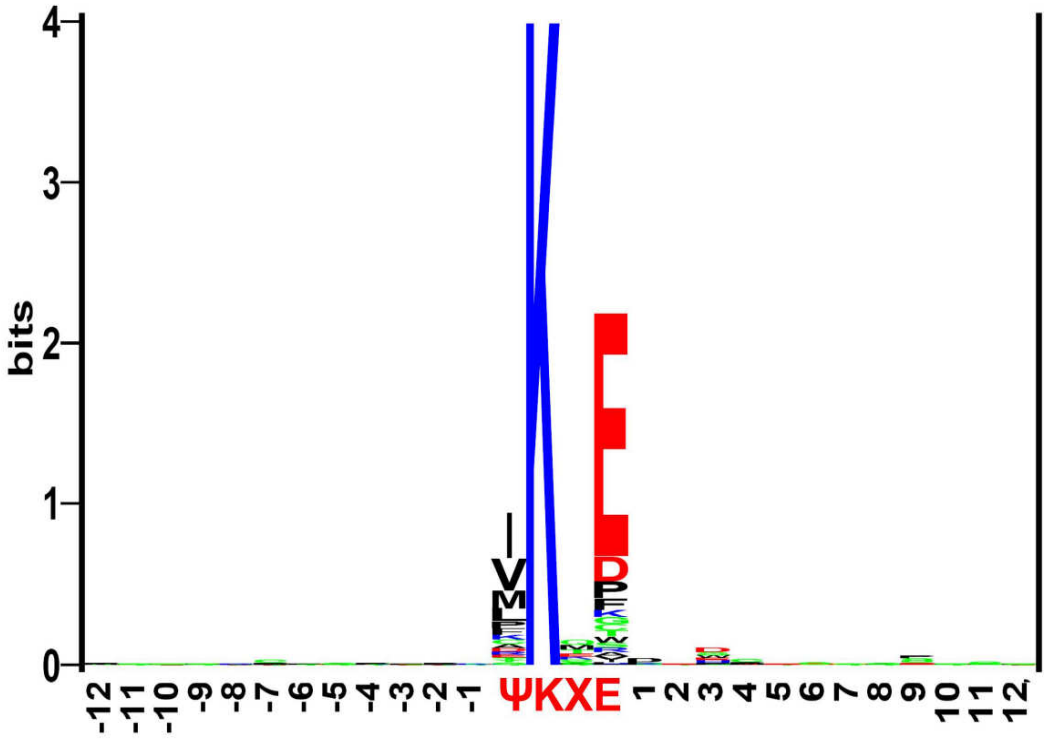

Figure 3.1 The sequence logo of the protein sumoylation motif (ΨKXE) and its flanking residues.

Effect of sequence context on classifier performance

We first constructed Random Forest (RF) classifiers using the forty biological features for input vector encoding. The RF classifiers were trained with data instances of various window sizes. The results suggest that RF classifier performance was affected by the sequence context of sumoylation sites (Table 3.1). The classifier constructed with KX (window size w = 2) gave predictive performance with the prediction strength (ST) = 57.07%, Matthews Correlation Coefficient (MCC) = 0.0590 and ROC AUC = 0.6107. The classifier performance was improved significantly when the core motif ΨKXE (w =

4) was used for input encoding. The classifier gave the prediction strength at 82.04% with MCC = 0.5379 and AUC = 0.9024. When the neighboring residues of the core motif were used to construct the classifiers, the predictive performance was further improved. For example, the classifier constructed with ΨKXE+/-5 (w = 14) achieved the highest MCC at 0.6786. Since the dataset was imbalanced with only 4% of lysine residues as the sumoylation sites, the ROC AUC is probably the most reliable performance measure for the present study. The classifier using the twenty residues with the core motif ΨKXE in the middle (ΨKXE+/-8, w = 20) reached the highest ROC AUC at 0.9200. The classifier also shows the highest overall accuracy at 97.68% with 56.00% sensitivity and 99.50% specificity, and high MCC = 0.6711. Thus, this RF classifier is considered as the best classifier in Table 3.1.

Table 3.1 Effect of sequence context on predictive performance of Random Forest classifiers.

Sequence context	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
KX	61.80	51.89	62.25	57.07	0.0590	0.6107
ΨKXE	95.27	67.57	96.52	82.04	0.5379	0.9024
ΨKXE+/-1	97.28	61.35	98.90	80.13	0.6489	0.9038
ΨKXE+/-2	97.42	59.73	99.12	79.42	0.6582	0.9145
ΨKXE+/-3	97.45	60.00	99.15	79.58	0.6638	0.9172
ΨKXE+/-4	97.54	60.28	99.20	79.74	0.6688	0.9074
ΨKXE+/-5	97.63	60.00	99.31	79.65	0.6786	0.9103
ΨKXE+/-6	97.57	57.78	99.34	78.56	0.6668	0.9048
ΨKXE+/-7	97.54	54.86	99.40	77.13	0.6508	0.9188
ΨKXE+/-8	97.68	56.00	99.50	77.75	0.6711	0.9200
ΨKXE+/-9	97.59	51.71	99.60	75.66	0.6522	0.9133
ΨKXE+/-10	97.57	47.65	99.70	73.67	0.6340	0.9149
ΨKXE+/-11	97.46	43.82	99.75	71.79	0.6115	0.9136
ΨKXE+/-12	97.43	42.35	99.79	71.07	0.6056	0.9124

The ROC analysis for investigating the effect of sequence context information on RF classifier performance has been shown in Figure 3.2. The classifier constructed with ΨKXE is clearly better than the classifier constructed with KX. Furthermore, the classifier using twenty residues with the core motif in the middle (ΨKXE+/-8) appears to be slightly better than the classifier constructed with ΨKXE. The results suggest that the context information in the flanking sequences may be useful for sumoylation site prediction.

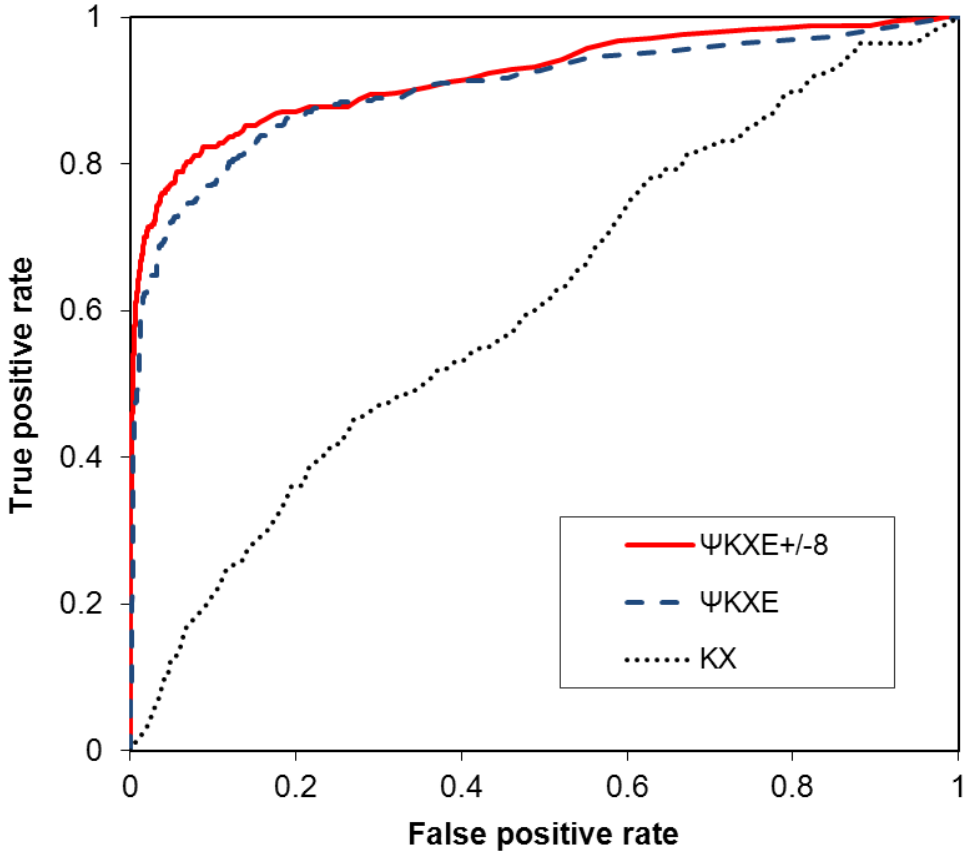

Figure 3.2 ROC curves to show the effect of context information for sumoylation site prediction.

RF versus SVM classifiers

Support Vector Machines (SVMs) have been widely used for biological pattern classification. In this study, we constructed SVM classifiers using the forty biological features, and compared their ROC AUC values with those of the RF classifiers over various window sizes. As shown in Figure 3.3, the RF classifiers using the forty features (RF40) achieved comparable performance measures over various window sizes with the highest AUC at w = 20 (ΨKXE+/-8). However, SVM classifiers using the forty features (SVM40) showed significantly degraded performances with large window sizes. For example, the AUC value of SVM40 decreased from 0.8090 to 0.5254 when the window size was increased from w = 4 (ΨKXE) to w = 8 (ΨKXE+/-2). Thus, the SVM classifiers did not achieve the same level of predictive performance as the RF classifiers. The possible explanation is that some of the forty features may contain redundant or correlated information for sumoylation site prediction, which may have caused the degradation of SVM classifier performance.

To enhance the predictive performance of SVM classifiers, feature selection was performed using Random Forests (RFs). Five highly relevant features selected by RFs, including polarity (P), conformational parameters for alpha helix (A) and coil (C), short and medium range non-bonded energy per residue (Er) and free energy in alpha-helical conformation (Ea), were used to construct the SVM classifiers (SVM5). As shown in Figure 3.3, the ROC AUC values of the SVM5 classifiers were higher than those of the SVM40 classifiers over various window sizes. For example, the SVM5 classifier constructed using eight residues (ΨKXE+/-2, w = 8) achieved the highest AUC value of

0.8917 over various window sizes (Figure 3.3), which was significantly higher than the AUC value of the SVM40 classifier (AUC = 0.5254 at *w* = 8). This classifier reached the prediction strength at 78.42% (58.38% sensitivity and 98.46% specificity) and MCC = 0.5902 (Table 3.2). The SVM5 classifier constructed with ΨKXE+/-2 was regarded as the best SVM classifier in this study.

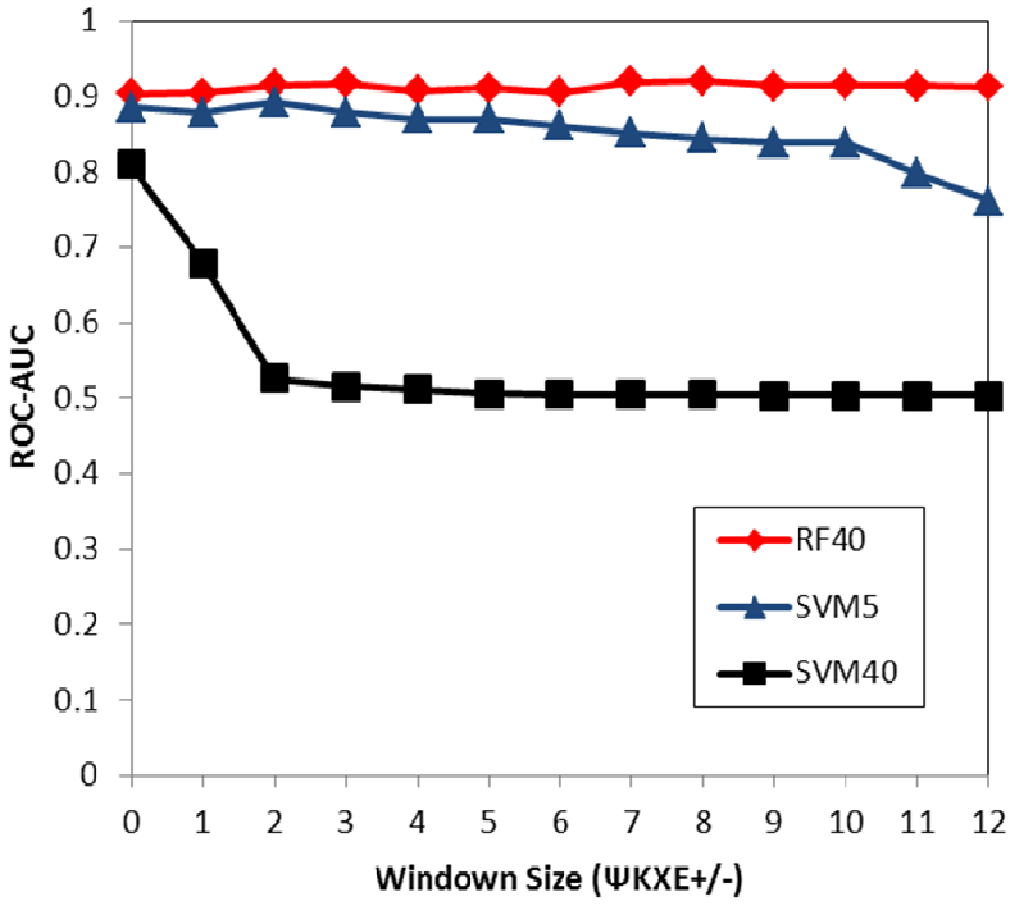

Figure 3.3 Performance comparisons of Random Forest (RF) and Support Vector Machine (SVM) classifiers.

However, the RF40 classifiers still outperformed the SVM5 models (Figure 3.3). As shown in Table 3.2, The RF40 classifier constructed using eight residues (ΨKXE+/-2) achieved the prediction strength of 79.42% with MCC = 0.6582 and AUC = 0.9145, which were higher than those of the SVM5 classifier in the same window size. Thus, the RF algorithm appears to be better for predicting protein sumoylation sites from sequence features. The possible explanation is that RFs can handle a large number of input variables and avoid model overfitting. The feature-encoded input vector has a large number of variables, especially with a large window size. For example, when twenty residues (w = 20) are used for classifier construction, the number of input variables is 800 for classifiers using forty features and 100 for classifiers using five features. The large number of input variables, together with the small number of positive instances, may lead to model overfitting.

Table 3.2 Comparison of Random Forest and Support Vector Machine classifiers constructed with ΨKXE+/-2 (w = 8).

Features	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
RF40	97.42	59.73	99.12	79.42	0.6582	0.9145
SVM5	96.73	58.38	98.46	78.42	0.5902	0.8917
SVM40	95.66	0.00	99.99	49.99	-0.0023	0.5254

Use of evolutionary information

Evolutionary information in terms of position-specific scoring matrix (PSSM) scores was previously shown to improve classifier performance [21, 25, 26]. To determine whether or not sumoylation site prediction could be further improved by using

evolutionary information, the PSSM scores of twenty residues with the core motif ΨKXE in the middle were used to construct the RF classifiers. The scores in a PSSM represent how well each position of a protein sequence was conserved among its homologues. As shown in Table 3.3, the RF classifier constructed with PSSMs (PSSM, Table 3.3) reached the prediction strength of 51.96% with MCC = 0.1566 and AUC = 0.8672. By using both PSSMs and the forty biological features for input vector encoding, the RF classifier (Bio + PSSM, Table 3.3) gave a relatively high classifier performance (74.82% prediction strength, MCC = 0.6443 and AUC = 0.9181). However, these performance measures were not significantly different from those of the RF classifier using the biological features only (Bio, Table 3.3).

Table 3.3 Effect of evolutionary information on protein sumoylation site prediction.

Features	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
PSSM	95.90	4.00	99.91	51.96	0.1566	0.8672
Bio	97.68	56.00	99.50	77.75	0.6711	0.9200
Bio + PSSM	97.56	50.00	99.64	74.82	0.6443	0.9181

The ROC curves of the three RF classifiers are compared in Figure 3.4. The results confirm that classifier performance is not improved by adding the evolutionary information to the biological features for input encoding. The possible explanation is that the PSSM, which is designed for PSI-BLAST searches, may not capture the evolutionary information for sumoylation site prediction. Another possibility is that the forty biological features may already contain the evolutionary information necessary for predicting protein sumoylation sites.

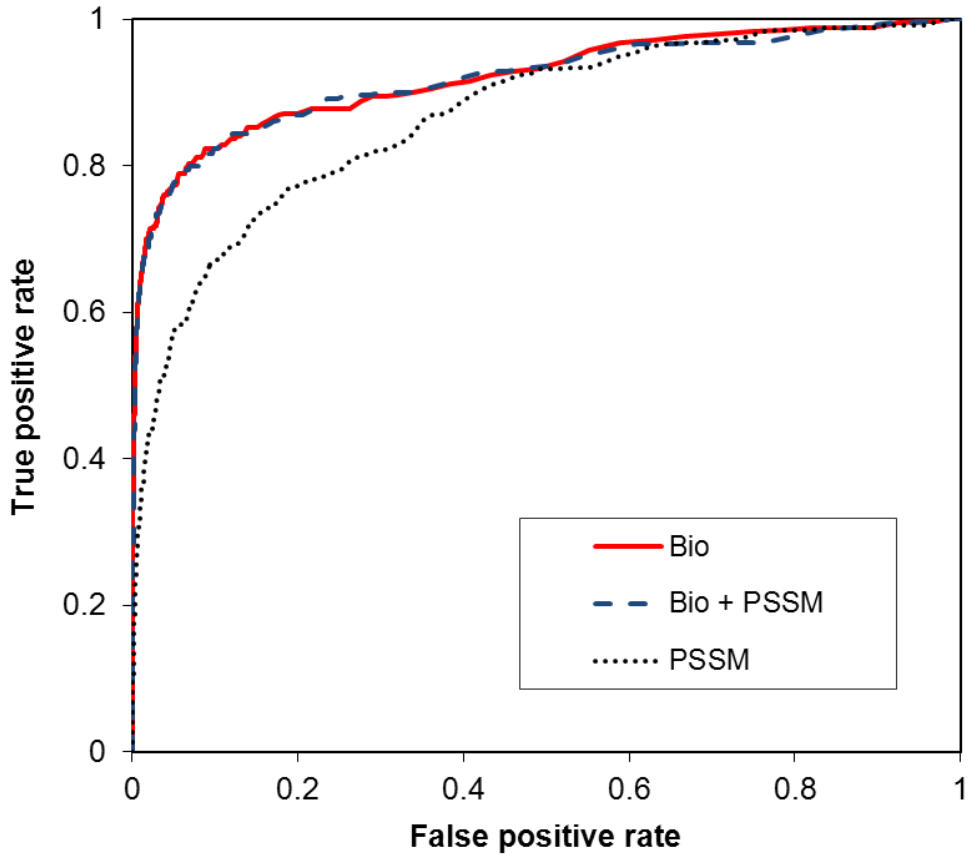

Figure 3.4 ROC curves to show the effect of evolutionary information on classifier performance.

Comparison with previous studies

The existing computational methods for protein sumoylation site prediction include SUMOplot (http://www.abgent.com//tools/toSumoplot), SUMOsp2 (http://sumosp.biocuckoo.org/online.php) and SUMOpre [9]. The datasets used in these previous studies are smaller than the dataset used in the present work. We manually collected additional instances of experimentally identified sumoylation sites from the latest publications. To further demonstrate the improved performance of our classifiers,

the most accurate RF classifier (ΨKXE+/-8, Table 3.1) and SVM classifier (SVM5, Table 3.2) have been compared with the previous classifiers, SUMOplot and SUMOsp2, using an independent test dataset with 48 sumoylation sites reported after January 2010. SUMOplot predicts the probability of sumoylation sites based on the SUMO consensus sequence and hydrophobicity, whereas SUMOsp2 [11] uses two pattern recognition strategies (GPS and MotifX) for sumoylation site prediction. The two types of prediction in SUMOplot were low (motifs with low probability) and high (motifs with high probability), whereas the three levels of stringency in SUMOsp2 were low, medium and high. The corresponding thresholds of classifier output in our approach (seeSUMO) were set to -0.2 (low), 0 (medium) and 0.2 (high).

As shown in Table 3.4, the overall accuracy (AC), specificity (SP) and MCC of our SVM classifier (seeSUMO-SVM) and RF classifier (seeSUMO-RF) are considerably higher than those of SUMOsp2 and SUMOplot in the low-threshold predictions. SUMOsp2 with its medium threshold gave the prediction strength at 68.31% (43.75% sensitivity and 92.87% specificity) and MCC = 0.2449. Our SVM classifier achieved a similar level of performance with 67.36% prediction strength, 41.67% sensitivity, 93.06% specificity and MCC = 0.2361. The RF classifier with the medium threshold reached higher performance with 71.60% prediction strength, 51.16% sensitivity, 92.04% specificity and MCC = 0.2639. For the high-threshold predictions, the overall accuracy and MCC of our RF classifier are also higher than those of SUMOsp2 and SUMOplot. It is noteworthy that the MCC values of our RF classifier are the highest in any level of threshold predictions. Therefore, the performance of the RF classifier developed in this

study compares favorably with SUMOsp2 and SUMOplot for protein sumoylation site prediction.

SUMOpre [9] uses a statistical method for predicting protein sumoylation sites. It was not included in the direct comparisons because a web-based tool was not available for the classifier. However, our classifier uses a larger dataset and shows better predictive performance. For example, the dataset used by SUMOpre [9] contains 268 sumoylation sites, and the predictor reached 96.71% overall accuracy and MCC = 0.6364. In the present study, the dataset includes 377 sumoylation sites, and the best RF classifier (ΨKXE+/-8, Table 3.1) achieved 97.68% overall accuracy and MCC = 0.6711

Table 3.4 Comparison of classifier performance using an independent test dataset.

Threshold	**Methods**	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**
Low	**SUMOplot**	79.55	68.75	79.95	74.35	0.2196
	SUMOsp2	83.63	50.00	84.88	67.44	0.1753
	seeSUMO-SVM	89.48	47.92	91.04	69.48	0.2382
	seeSUMO-RF	90.09	53.49	91.33	72.41	0.2644
Medium	**SUMOsp2**	91.11	43.75	92.87	68.31	0.2449
	seeSUMO-SVM	91.21	41.67	93.06	67.36	0.2361
	seeSUMO-RF	90.70	51.16	92.04	71.60	0.2639
High	**SUMOplot**	91.69	50.00	93.24	71.62	0.2916
	SUMOsp2	94.24	39.58	96.27	67.93	0.3058
	seeSUMO-SVM	92.49	39.58	94.47	67.02	0.2528
	seeSUMO-RF	94.36	44.19	96.06	70.12	0.3210

seeSUMO web server

To make our classifiers accessible to the biological research community, we have developed the seeSUMO web server (http://bioinfo.ggc.org/seesumo/). Users can enter an amino acid sequence in the FASTA format, specify the methods, and input the proper threshold for prediction of protein sumoylation site. For prediction using the RF classifier, the system encodes the input sequences with the 40 biological features, and then calls the randomForest program of the R software package to classify the protein sumoylation sites using the most accurate RF model (ΨKXE+/-8, Table 3.1). For prediction using the SVM classifier, the system encodes the input sequences with the five highly relevant features, and then the best SVM classifier (SVM5, Table 3.2) constructed in this work is used to predict sumoylation sites in the query sequence. The seeSUMO web server will return the prediction results, including the protein name, potential sumoylated sites, classifier outputs and the prediction confidence levers (Figure 3.5). The prediction confidence lever is calculated as (1 - sensitivity) for positive predictions, and (1 - specificity) for negative predictions [12, 14]. The help documents are available at the website.

.

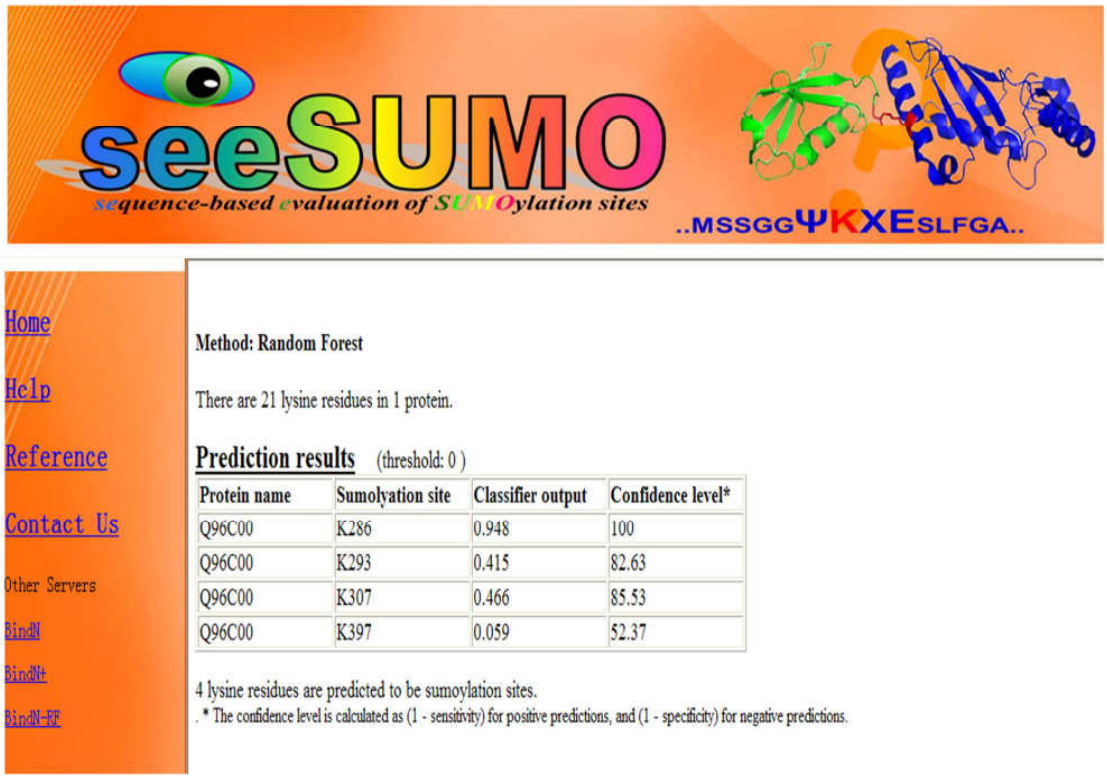

Figure 3.5 Sample output from the seeSUMO web server.

CONCLUSION

A new machine learning approach has been developed in this study for predicting protein sumoylation sites from protein sequence information. Domain-specific knowledge in terms of relevant biological features was used for input vector encoding. The results suggest that classifier performance is affected by the sequence context of sumoylation sites. The highest predictive performance (ROC AUC = 0.9200) has been achieved by the Random Forest classifier using twenty residues with the core motif ΨKXE in the middle. Moreover, the Random Forest classifiers were found to outperform

Support Vector Machine models on the imbalanced dataset. The classifiers developed in this study compare favourably in performance with the previous predictors for protein sumoylation site prediction. A web server, seeSUMO (http://bioinfo.ggc.org/seesumo/), has been developed to make our classifiers accessible to the biological research community.

REFERENCES

1. Zhao J: Sumoylation regulates diverse biological processes. *Cell Mol Life Sci* 2007, 64(23):3017-3033.

2. Geiss-Friedlander R, Melchior F: Concepts in sumoylation: a decade on. *Nat Rev Mol Cell Biol* 2007, 8(12):947-956.

3. Steffan JS, Agrawal N, Pallos J, Rockabrand E, Trotman LC, Slepko N, Illes K, Lukacsovich T, Zhu YZ, Cattaneo E, Pandolfi PP, Thompson LM, Marsh JL: SUMO modification of Huntingtin and Huntington's disease pathology. *Science* 2004, 304(5667):100-104.

4. Sarge KD, Park-Sarge OK: Sumoylation and human disease pathogenesis. *Trends Biochem Sci* 2009, 34(4):200-205.

5. Martin S, Wilkinson KA, Nishimune A, Henley JM: Emerging extranuclear roles of protein SUMOylation in neuronal function and dysfunction. *Nat Rev Neurosci* 2007, 8(12):948-959.

6. Yang SH, Galanis A, Witty J, Sharrocks AD: An extended consensus motif enhances the specificity of substrate modification by SUMO. *Embo J* 2006, 25(21):5083-5093.

7. Hietakangas V, Anckar J, Blomster HA, Fujimoto M, Palvimo JJ, Nakai A, Sistonen L: PDSM, a motif for phosphorylation-dependent SUMO modification. *Proc Natl Acad Sci U S A* 2006, 103(1):45-50.

8. Stankovic-Valentin N, Deltour S, Seeler J, Pinte S, Vergoten G, Guerardel C, Dejean A, Leprince D: An acetylation/deacetylation-SUMOylation switch through a phylogenetically conserved psiKXEP motif in the tumor suppressor HIC1 regulates transcriptional repression activity. *Mol Cell Biol* 2007, 27(7):2661-2675.

9. Xu J, He Y, Qiang B, Yuan J, Peng X, Pan XM: A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics* 2008, 9:8.

10. Xue Y, Zhou F, Fu C, Xu Y, Yao X: SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* 2006, 34(Web Server issue):W254-257.

11. Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y: Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics* 2009, 9(12):3409-3412.

12. Teng S, Srivastava AK, Wang L: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* 2010, 11 Suppl 2:S5.

13. Wang L, Brown SJ: Prediction of RNA-binding residues in protein sequences using support vector machines. *Conf Proc IEEE Eng Med Biol Soc* 2006, 1:5830-5833.

14. Wang L, Brown SJ: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006, 34(Web Server issue):W243-248.

15. Ahmad S, Gromiha MM, Sarai A: Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 2004, 20(4):477-486.

16. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, 23(21):2947-2948.

17. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, 18(20):6097-6100.

18. Gorodkin J, Heyer LJ, Brunak S, Stormo GD: Displaying the information contents of structural RNA alignments: the structure logos. *Comput Appl Biosci* 1997, 13(6):583-586.

19. Gasteiger E. HC, Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.: The Proteomics Protocols Handbook: Humana Press; 2005.

20. Kawashima S, Kanehisa M: AAindex: amino acid index database. *Nucleic Acids Res* 2000, 28(1):374.

21. Ahmad S, Sarai A: PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005, 6:33.

22. Noble WS: What is a support vector machine? *Nat Biotechnol* 2006, 24(12):1565-1567.

23. Swets JA: Measuring the accuracy of diagnostic systems. *Science* 1988, 240(4857):1285-1293.

24. Bradley A: The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997, 30:1145-1159.

25. Wang L, Huang C, Yang MQ, Yang JY: BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010, 4 Suppl 1:S3.

26. Pu X, Guo J, Leung H, Lin Y: Prediction of membrane protein types from sequences and position-specific scoring matrices. *J Theor Biol* 2007, 247(2):259-265.

CHAPTER FOUR

SEQUENCE FEATURE-BASED PREDICTION OF PROTEIN STABILITY CHANGES UPON AMINO ACID SUBSTITUTIONS[3]

ABSTRACT

Protein destabilization is a common mechanism by which amino acid substitutions cause human diseases. Although several machine learning methods were reported for predicting protein stability changes upon amino acid substitutions, these previous studies did not utilize relevant sequence features representing biological knowledge for classifier construction. In this study, a new machine learning method has been developed for predicting protein stability changes upon amino acid substitutions from sequence features. Support vector machines were trained with data from experimental studies on the free energy change of protein stability upon mutations. To construct accurate classifiers, twenty sequence features were examined for input vector encoding. It was shown that classifier performance varied significantly by using different features. The most accurate classifier in this study was constructed using a combination of six sequence features. The classifier achieved an overall accuracy of 84.59% with 70.29% sensitivity and 90.98% specificity. Protein stability changes upon amino acid substitutions can be predicted accurately from relevant sequence features. Since predictive results at this level of accuracy may provide useful information to distinguish between deleterious and tolerant alterations in disease candidate genes, we have

[3]Teng S, Srivastava AK, Wang L: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 2010, 11(Suppl 2):S5.

developed a new web server, called MuStab (http://bioinfo.ggc.org/mustab/), to make the classifier accessible to the genetics research community.

BACKGROUND

Amino acid substitutions may cause a series of changes to normal protein function, such as geometric constraint changes, physico-chemical effects, and disruption of salt bridges or hydrogen bonds [1]. These changes may lead to protein destabilization or some abnormal biological functions. Previous studies suggest that each person may have 24,000 – 40,000 non-synonymous Single Nucleotide Polymorphisms (nsSNPs), and there are a total of 67,000 – 200,000 common nsSNPs in the human population [2]. These nsSNPs give rise to amino acid substitutions in proteins. While most nsSNPs appear to be functionally neutral, the others affect protein function and may cause diseases. Yue and Moult [3] investigated the effect of amino acid substitutions on protein stability, and estimated that approximately 25% of nsSNPs in the human population might be deleterious to protein function. Of the known disease-causing missense mutations, the vast majority (up to 80%) resulted in protein destabilization [4]. However, it is not feasible to experimentally determine the effect of each human nsSNP on protein stability. Rather, computational methods are needed to provide fast and efficient tools for examining a large number of nsSNPs for potential disease-causing mutations.

Machine learning has recently been applied to sequence-based prediction of protein stability changes upon amino acid substitutions [5]. The machine learning problem can be specified as follows: given the amino acid sequence of a protein and a

single amino acid substitution, the task is to predict whether the substitution may alter protein stability. By using the available data from experimental studies, classifiers can be constructed for predicting either the free energy change ($\Delta\Delta G$) of protein stability upon mutations or the direction of the change (increased stability if $\Delta\Delta G > 0$, or decreased stability if $\Delta\Delta G < 0$). Nevertheless, for many biological applications, correctly predicting the direction of the stability change (a binary classification problem) is more relevant than estimating the magnitude of the free energy change (a regression problem) [5].

Capriotti et al. [5] reported an artificial neural network-based method for predicting the direction of protein stability changes upon point mutations. The predictor was trained with protein sequence alone. It was shown that the sequence-based system could be used to complement the available energy-based methods for improving protein design strategies. The same research group also developed support vector machine (SVM) models for sequence-based prediction of both the free energy change and the direction of the change upon mutations [6]. These SVM models were used to develop the I-Mutant2.0 web server, which could predict protein stability increase or decrease at the overall accuracy of 77% (based on cross-validation). Interestingly, it was found that the sequence-based system was almost as accurate as the structure-based method (80% overall accuracy) on the same dataset [6]. This observation was further confirmed by Cheng et al., who trained SVMs for predicting protein stability changes from amino acid sequence and structural information [7]. More recently, Huang et al. [8] developed the iPTREE-STAB web server, which used decision trees with an adaptive boosting algorithm to discriminate stabilizing and destabilizing substitutions in protein sequences.

Among all the existing methods, iPTREE-STAB achieved the best classifier performance in cross-validation tests (82.1% overall accuracy with 75.3% sensitivity and 84.5% specificity).

The above-mentioned studies suggest that protein stability changes can be predicted directly from primary sequence data with similar prediction accuracy as structure-based methods. The sequence-based approach is particularly appealing since structural information is still not available for most proteins. However, little domain-specific knowledge in terms of biological features was used for classifier construction in the previous studies [5]. In the present study, we have examined twenty sequence features for classifier construction. Support vector machines (SVMs) have been trained with the feature-encoded data instances of protein stability changes upon amino acid substitutions. Our results indicate that accurate SVM classifiers can be constructed using relevant sequence features for input vector encoding.

METHODS

Data

The dataset used in this study was derived from two previous studies [6, 8], in which experimental data for the free energy changes of protein stability upon mutations were collected from the ProTherm database [9] . To construct a robust classifier, data redundancy was removed and the dataset had less than 25% identity among the amino acid sequences. Each data instance in the dataset had the following attributes: amino acid sequence, wide-type amino acid identity and sequence position, mutant amino acid

identity, pH value, and free energy change. If the free energy change was negative (protein destabilization), the instance was labelled as a negative example. Otherwise, the instance was labelled as a positive example. The dataset contained 464 positive instances and 1,016 negative instances.

Sequence features

Twenty sequence features were used to code each amino acid residue in a data instance. The sequence features were obtained from Protscale [10] (http://expasy.org/tools/protscale.html) and AAindex [11] (http://www.genome.jp/aaindex/). These features fall into the following four classes:

1) Biochemical features: including molecular weight (feature M); side-chain pKa value (K); hydrophobicity index (H); polarity (P); and overall amino acid composition (Co). Each amino acid has a unique molecular weight (M), which is related to the volume of space that a residue occupies in protein structures. Side-chain pKa (K) is related to the ionization state of a residue, and thus plays a key role in pH-dependent protein stability. Hydrophobicity (H) is important for amino acid side chain packing and protein folding. Hydrophobic interactions make non-polar side chains to pack together inside proteins, and disruption of these interactions may cause protein destabilization. Polarity (P) is the dipole-dipole intermolecular interactions between the positively and negatively charged residues. The amino acid composition (Co) was previously shown to be related to the evolution and stability of small proteins [12].

2) Structural features: including the conformational parameters for alpha-helix (A), beta-sheet (B), and coil (C); average area buried on transfer from standard state to folded protein (Aa); and bulkiness (Bu). Protein secondary structures can be divided into alpha-helix, beta-sheet, and coil conformations. An amino acid often has a different tendency to form one of the three types of secondary structures. For instance, amino acids A, I, E, L and M tend to be in the alpha-helical conformation, whereas K, N and D are often found in beta-sheets. In this study, the conformational parameters reported by Deléage and Roux [13] were used for features A, B and C. Feature Aa is another structural parameter, which estimates a residue's average area buried in the interior core of a globular protein [14]. Bulkiness (Bu), the ratio of the side chain volume to the length of an amino acid, may affect the local structure of a protein [15].

3) Empirical features: the protein stability scale based on atom-atom potential (S1); the relative protein stability scale derived from mutation experiments (S2); and the side-chain contribution to protein stability (S3). Zhou et al. [16] derived two protein stability scales from atom-atom potential of mean force based on Distance scaled Finite Ideal-gas REference (DFIRE) state (S1) and a large database of mutations (S2). Takano and Yutani [17] calculated the transfer Gibbs energy of mutant proteins, and derived the amino acid scale for the side-chain contribution to protein stability (S3) based on data from protein denaturation experiments.

4) Other biological features: including the average flexibility index (F); the mobility of an amino acid on chromatography paper (Mc); the number of codons for an amino acid (No); refractivity (R); recognition factor (Rf); the relative mutability of an amino acid

(Rm); and transmembrane tendency (Tt). The average flexibility index of an amino acid (F) was derived from structures of globular proteins [18]. Feature Mc was derived from experimental data by Aboderin [19]. Refractivity (R) refers to protein density and folding characteristics [20]. Recognition factor (Rf) is the average of stabilization energy for an amino acid [21]. The relative mutability (Rm) indicates the probability that a given amino acid can be changed to others during evolution. Feature Tt is the transmembrane tendency scale described by Zhao and London [22].

Support vector machine training

Support vector machines (SVMs) are computational algorithms that can learn from training examples for binary classification problems. The SVM learning algorithm can be described by four basic concepts, including the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function [23]. For a typical linear classifier, a data instance is represented as an n-dimensional vector, and an $(n - 1)$ dimensional hyperplane is used to separate the positive instances from the negative ones. However, for non-linear classifiers that are generally applicable to biological problems, a kernel function can be used to measure the distance between data points in a higher dimensional space. This allows the SVM algorithm to fit the maximum-margin hyperplane in the transformed space. In this study, we used the radial basis function (RBF) kernel:

$$K(\vec{x}, \vec{y}) = \exp(-\gamma \| \vec{x} - \vec{y} \|^2) \quad (4.1)$$

where $\vec{x}$ and $\vec{y}$ are two data vectors, and γ is a training parameter. A smaller γ value makes decision boundary smoother. The regularization factor C, another parameter for SVM training, controls the tradeoff between low training error and large margin.

The SVMlight software package (available at http://svmlight.joachims.org/) was used to construct the SVM classifiers in this study. Each training instance was a subsequence of *w* consecutive residues, where *w* was also called the window size. The amino acid substitution site was positioned in the middle of the subsequence, and the other ($w - 1$) neighbouring residues provided context information for the substitution site. The input vector was then obtained by encoding each residue with one or more biological features. The input vector also included the pH value at which the free energy change was measured experimentally. In this study, various values of *w*, γ and *C* parameters were examined to optimize SVM classifier performance.

Classifier evaluation

This study used a fivefold cross-validation method to evaluate classifier performance. Positive and negative instances were randomly distributed into five folds. In each of the five iterations, four of the five folds were used to train a classifier, and then the remaining one fold was used as the test data to evaluate the classifier. The predictions made for the test instances in all the five iterations were combined and used to compute the following performance measures:

$$\text{Accuracy (AC)} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4.2)$$

$$\text{Sensitivit y (SN)} = \frac{TP}{TP + FN} \tag{4.3}$$

$$\text{Specificiy (SP)} = \frac{TN}{TN + FP} \tag{4.4}$$

$$\text{Strength (ST)} = \frac{\text{Sensitivit y} + \text{Specificit y}}{2} \tag{4.5}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.6}$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; and FN is the number of false negatives. In addition to the commonly used performance measures (overall accuracy, sensitivity and specificity), the average of sensitivity and specificity or the so-called prediction strength [24, 25] was also used for classifier comparison in this study. Matthews Correlation Coefficient (MCC) measures the correlation between predictions and the actual class labels. Nevertheless, for imbalanced datasets, different tradeoffs of sensitivity and specificity may give rise to different MCC values for a classifier.

We also used the Receiver Operating Characteristic (ROC) curves [26] for classifier evaluation and comparison. In this study, the ROC curve was generated by varying the output threshold of an SVM classifier and plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) for each threshold value. Since the ROC curve of an accurate classifier is close to the left-hand and top borders of the plot, the area under the curve (AUC) can be used as a reliable measure of classifier performance [27]. The maximum value of AUC is 1, which indicates a perfect classifier. Weak classifiers and random guessing have AUC values close to 0.5.

RESULTS AND DISCUSSION

Effect of sequence context on classifier performance

We first constructed a classifier using the three biochemical features, including the hydrophobicity index (H), side-chain pKa value (K), and molecular weight (M) of an amino acid. These features were previously selected for DNA and RNA-binding site prediction [24, 25]. In the initial attempt to construct a classifier for protein stability prediction, the window size was set to eleven (w = 11). Different values of SVM training parameters were tested, and the optimal parameter settings were found to be γ = 0.8 and C = 1.0. As shown in Table 6.1, the classifier achieved the overall accuracy (AC) of 81.82% with 74.48% sensitivity (SN) and 85.11% specificity (SP). The prediction strength (ST) reached 79.79% with MCC = 0.5843 and ROC AUC = 0.8804. Therefore, this SVM achieved similar performance measures as the best existing classifier (iPTREE-STAB with 82.1% overall accuracy, 75.3% sensitivity and 84.5% specificity) [8].

To determine whether classifier performance was affected by the sequence context of the substitution site, SVMs were trained with data instances of various window sizes. As shown in Table 4.1, protein stability prediction was affected by window sizes. The classifier constructed without any context information (w = 1) gave 67.94% prediction strength (70.69% sensitivity and 65.20% specificity), MCC = 0.3349 and AUC = 0.7425. The prediction strength, MCC and AUC were improved when neighbouring residues of the substitution site were included for input encoding. The use of w = 11 gave the highest prediction strength (79.79%), MCC (0.5843) and AUC (0.8804), and

classifier performance was not further improved by including more neighbouring residues (Table 4.1).

Table 4.1 Effect of window sizes on sequence-based prediction of protein stability changes.

Window size	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
1	66.92	70.69	65.20	67.94	0.3349	0.7425
3	73.91	74.83	73.49	74.16	0.4554	0.7996
5	77.51	76.67	77.90	77.28	0.5194	0.8512
7	80.80	76.43	82.83	79.63	0.5750	0.8737
9	81.28	75.66	83.78	79.72	0.5774	0.8755
11	81.82	74.48	85.11	79.79	0.5843	0.8804
13	82.10	71.84	86.67	79.26	0.5824	0.8797
15	81.45	69.71	86.75	78.23	0.5665	0.8775
17	81.88	69.50	87.58	78.54	0.5779	0.8799
19	81.21	68.80	86.98	77.89	0.5627	0.8779
21	81.29	68.98	86.98	77.98	0.5645	0.8735

The effect of sequence context information on SVM classifier performance was also demonstrated by using ROC curves. As shown in Figure 4.1, the ROC curve of the classifier constructed with $w = 11$ was clearly better than the SVM trained without any context information ($w = 1$). However, the use of $w = 21$ did not further improve classifier performance. Thus, eleven residues with the substitution site in the middle position ($w = 11$) appeared to provide enough context information for sequence-based prediction of protein stability changes.

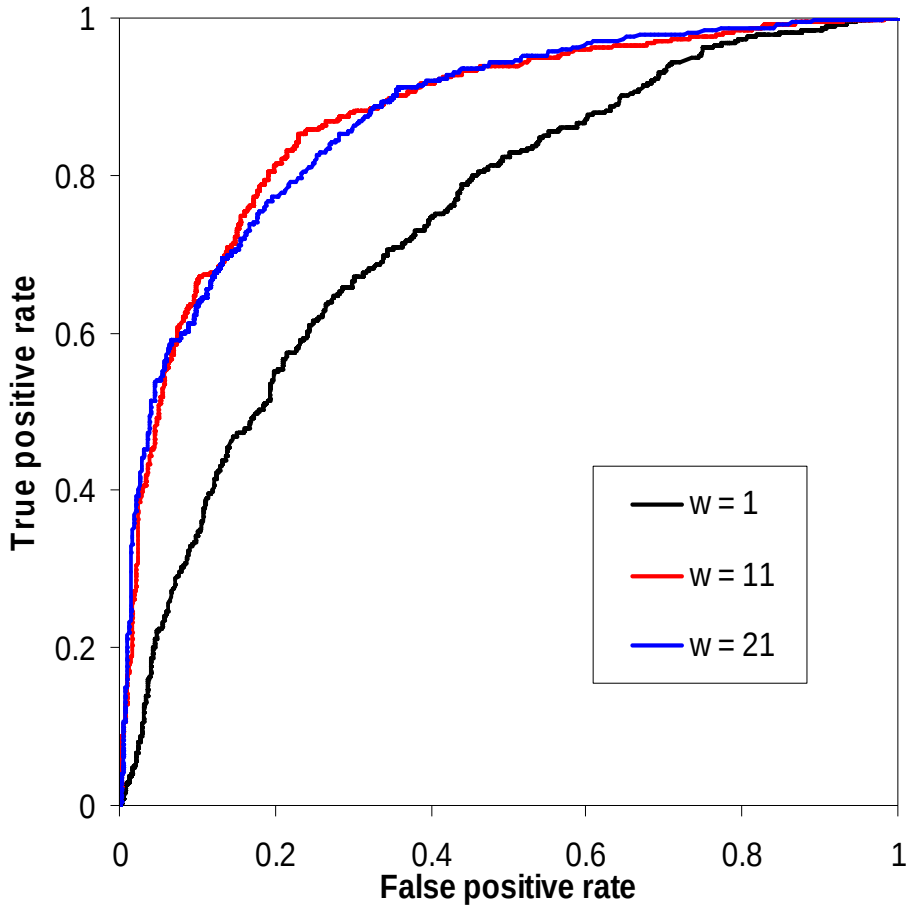

Figure 4.1 ROC curves to show the effect of context information on prediction of protein stability changes upon amino acid substitutions.

Relevant sequence features for classifier construction

Many sequence features are available for encoding amino acid residues. To determine which features were relevant for protein stability prediction, we constructed SVM classifiers using each of the twenty sequence features listed in Table 4.2 for input encoding (w = 11). The results were obtained with the training parameters, γ = 0.8 and C = 1.0. It was found that classifier performance varied significantly by using different features. As shown in Table 6.2, the highest level of AUC (0.8835) was achieved by

using the empirical feature S3 for input encoding. This classifier reached the prediction strength at 79.67% (72.19% sensitivity and 87.15% specificity) and MCC = 0.5922. However, the highest prediction strength at 80.28% (75.62% sensitivity and 84.94% specificity) with MCC = 0.5919 and AUC = 0.8777 was achieved by using amino acid bulkiness (Bu) for input encoding. In contrast, the use of the average flexibility index (F) for input encoding resulted in the lowest prediction strength at 62.02%, MCC = 0.2226 and AUC = 0.6728 (Table 4.2).

Table 4.2 Predictive performance of classifiers constructed using single sequence features.

Features	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
H	75.88	71.62	77.79	74.70	0.4728	0.8237
K	73.29	73.90	73.02	73.46	0.4402	0.7925
M	68.06	73.52	65.62	69.57	0.3629	0.7480
P	75.94	71.24	78.04	74.64	0.4718	0.8234
Co	70.18	71.62	69.53	70.58	0.3838	0.7586
A	76.41	74.29	77.36	75.82	0.4904	0.8206
B	78.18	74.48	79.83	77.15	0.5199	0.8503
C	72.18	71.05	72.68	71.86	0.4116	0.7847
Aa	79.12	76.57	80.26	78.41	0.5431	0.8459
Bu	82.06	75.62	84.94	80.28	0.5919	0.8777
S1	69.82	70.86	69.36	70.11	0.3756	0.7754
S2	70.24	72.19	69.36	70.78	0.3875	0.7665
S3	82.53	72.19	87.15	79.67	0.5922	0.8835
F	61.41	63.62	60.43	62.02	0.2226	0.6728
R	66.47	65.14	67.06	66.10	0.3008	0.7140
Mc	78.35	73.52	80.51	77.02	0.5202	0.8417
No	69.82	74.86	67.57	71.22	0.3944	0.7656
Rf	62.06	73.71	56.85	65.28	0.2831	0.6889
Rm	75.94	69.90	78.64	74.27	0.4672	0.8118
Tt	83.59	66.48	91.23	78.86	0.6035	0.8704

Figure 4.2 shows the ROC curves of the best and worst classifiers (based on AUC) that were constructed using the individual sequence features. Also shown in Figure 6.2 is the ROC curve of the SVM classifier constructed with the K feature, which gave approximately the average performance among the sequence features.

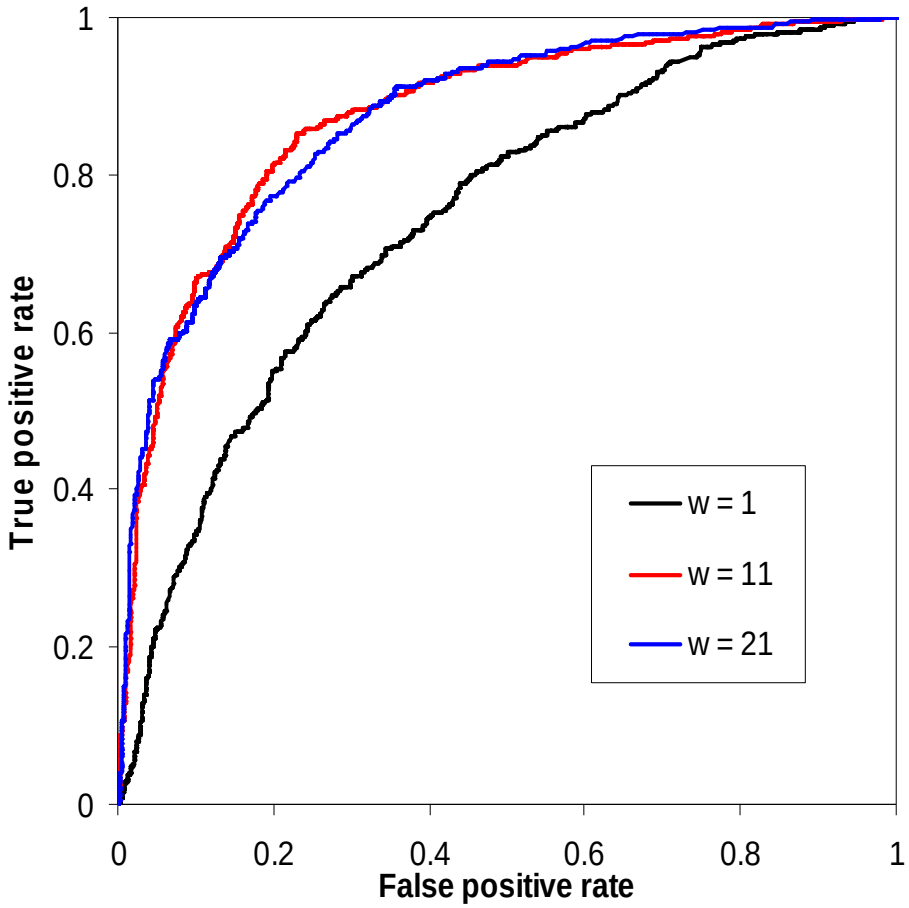

Figure 4.2 ROC curves to show the different performance levels of classifiers constructed using individual sequence features.

The results suggest that a variety of sequence features are relevant for predicting protein stability changes upon amino acid substitutions. Of the five biochemical features

(H, K, M, P and Co), the hydrophobicity index (H) gave the best predictive performance at 74.70% prediction strength (71.62% sensitivity and 77.79% specificity), MCC = 0.4728 and AUC = 0.8237 (Table 4.2). Hydrophobicity is a key factor in amino acid side chain packing and protein folding. Hydrophobicity changes owing to amino acid substitutions may cause proteins not to fold into stable conformation, and thus result in protein destabilization.

Of the structural features (A, B, C, Aa and Bu), bulkiness (Bu) gave rise to the highest prediction strength at 80.28% with MCC = 0.5919 and AUC = 0.8777. In contrast, the classifier using the conformational parameter for coil (C) had the relatively low performance with 71.86% prediction strength, MCC = 0.4116 and AUC = 0.7847 (Table 4.2). The possible explanation is that since coils are often unstructured and flexible, amino acid substitutions in the coil region may not cause significant changes in protein structure and stability.

The empirical features (S1, S2 and S3) are protein stability scales based on experimental data. Interestingly, when used for SVM classifier construction, these features did not give significantly better performance than the other sequence features. While the use of the S3 feature (side-chain contribution to protein stability) resulted in the highest level of AUC (0.8835) with 79.67% prediction strength and MCC = 0.5922, the other two empirical features (S1 and S2) were much less accurate for predicting protein stability changes (Table 4.2). Thus, it is possible that the empirical features do not capture all the information about the determinants of protein stability.

Of the other biological features, transmembrane tendency (Tt) achieved the highest level of MCC (0.6035) with 78.86% prediction strength and AUC = 0.8704 (Table 4.2). The feature Mc (the mobility of an amino acid on chromatography paper) also gave rise to relatively high classifier performance (77.02% prediction strength, MCC = 0.5202 and AUC = 0.8417). Therefore, multiple features from each of the four feature classes achieved high performance for predicting protein stability changes upon amino acid substitutions. It might be possible that classifier performance could be further improved by combining several sequence features for input encoding.

Use of multiple sequence features to improve classifier performance

To examine whether classifier performance could be further improved, we first used all the 20 sequence features for input encoding. Surprisingly, the resulting classifier was not as accurate as some of the SVMs trained with single features (Table 4.3). While the best single feature S3 gave rise to 79.67% prediction strength with MCC = 0.5922 and AUC = 0.8835, the classifier using all the 20 features achieved only 75.45% prediction strength with MCC = 0.5791 and AUC = 0.8690. The possible explanation is that some of the 20 features contain redundant or correlated information, which may cause classifier performance degradation.

We then constructed SVM classifiers by combining some of the best single features for input encoding. Interestingly, none of these feature combinations gave rise to better classifier performance than the best single feature S3 (Table 4.3). For example, the

classifier constructed using the best six single features (S3, Bu, Tt, B, Aa, and Mc) achieved only 77.54% prediction strength with MCC = 0.5993 and AUC = 0.8737.

Table 4.3 Predictive performance of classifiers constructed by combining the best single features.

Features	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
S3	82.53	72.19	87.15	79.67	0.5922	0.8835
S3, Bu	83.41	68.00	90.30	79.15	0.6019	0.8821
S3, Bu, Tt	82.88	61.90	92.26	77.08	0.5822	0.8725
S3, Bu, Tt, B	83.65	62.10	93.28	77.69	0.6009	0.8768
S3, Bu, Tt, B, Aa	83.65	61.90	93.36	77.63	0.6009	0.8743
S3, Bu, Tt, B, Aa, Mc	83.59	61.71	93.36	77.54	0.5993	0.8737
All 20 features	82.88	56.00	94.89	75.45	0.5791	0.8690

To determine whether any combinations of the sequence features could improve classifier performance, we performed a brute-force search for the optimal feature subset. As shown in Table 4.4, classifier performance based on AUC was improved slightly but steadily when more features were used for input encoding. Among all the two-feature combinations, the biochemical feature Co (overall amino acid composition) together with the structural feature Bu (bulkiness) achieved the best classifier performance based on AUC (0.8872) with 80.54% prediction strength and MCC = 0.6057. These performance measures are slightly better than those of the empirical feature S3, a protein stability scale based on experimental data [17]. Significantly, the feature Co is also included in all the other feature subsets shown in Table 4.4, suggesting that the overall amino acid composition is highly relevant for sequence-based prediction of protein stability changes. For instance, the best four-feature subset contains the biochemical features Co and H

(hydrophobicity index), the structural feature B (conformational parameter for beta-sheet), and the empirical feature S3. The classifier achieved 80.16% prediction strength with MCC = 0.6231 and AUC = 0.8940 (Table 4.4).

Table 4.4 Predictive performance of classifiers constructed using the optimal subsets of sequence features.

Features	**AC (%)**	**SN (%)**	**SP (%)**	**ST (%)**	**MCC**	**ROC AUC**
S3	82.53	72.19	87.15	79.67	0.5922	0.8835
Bu, Co	83.00	74.10	86.98	80.54	0.6057	0.8872
B, Co, S3	84.12	69.33	90.72	80.03	0.6194	0.8924
B, Co, H, S3	84.29	69.33	90.98	80.16	0.6231	0.8940
A, Aa, B, Co, P	84.47	70.48	90.72	80.60	0.6287	0.8954
A, Aa, B, Co, No, P	84.59	70.29	90.98	80.63	0.6310	0.8961

As shown in Table 4.4, the highest performance measures were obtained by using the optimal subset of six features, including the biochemical features Co and P (polarity), the structural features A (conformational parameter for alpha-helix), B and Aa (average area buried on transfer from standard state to folded protein), and the other biological feature No (number of codons for an amino acid). Classifier performance was not further improved significantly by including additional sequence features (data not shown). Interestingly, the optimal feature subset did not include the best single feature S3. The classifier constructed using the optimal feature subset achieved 80.63% prediction strength with MCC = 0.6310 and AUC = 0.8961. In Figure 6.3, this classifier's ROC curve is compared with those of two other classifiers, one constructed using the best single feature S3, and the other trained with all the 20 features.

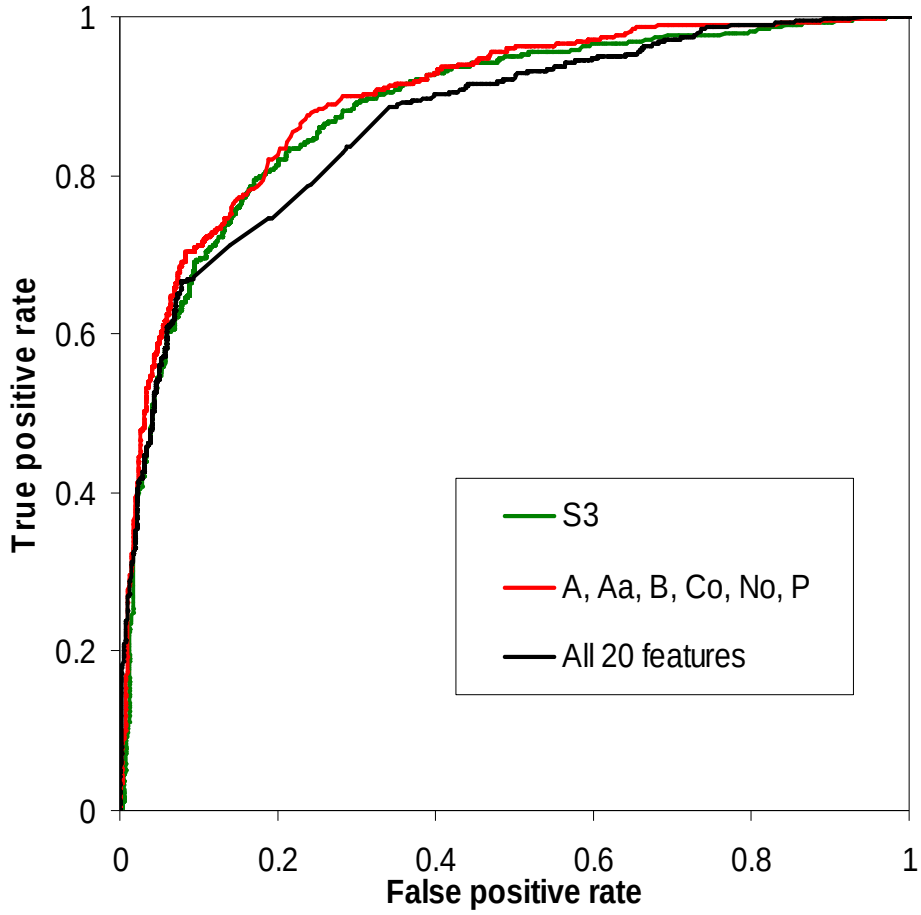

Figure 4.3 ROC curves for sequence-based prediction of protein stability changes using multiple sequence features.

The results suggest that classifier performance can be enhanced by combining certain sequence features for input encoding. The optimal six-feature subset contains sequence features from different classes, especially biochemical features and structural features. Each of these features may not be an accurate scale of protein stability, but when combined, they can outperform the best empirical feature (S3) for predicting protein stability changes upon amino acid substitutions.

Web server description

To make the accurate SVM classifier accessible to the biological research community, we have developed the MuStab web server (http://bioinfo.ggc.org/mustab/). Users can enter an amino acid sequence in FASTA format, and specify the position and the identity of the substituting residue. The system encodes the input sequence with the optimal feature subset, and then calls the svm_classify program of the SVMlight software package to classify the protein stability changes upon the amino acid substitution using the best SVM model developed in this study.

The output report returned from the MuStab web server includes the information about the query sequence and amino acid substitution, the prediction result, and the prediction confidence. The prediction result indicates either decreased or increased protein stability. The prediction confidence is based on the SVM output and computed as (1 – s), where s is the expected sensitivity for positive predictions or the expected specificity for negative predictions if the SVM output is used as the threshold in the ROC analysis (Figure 4.3). An example output report returned from the MuStab web server is shown in Figure 4.4 for the G56S substitution of spermine synthase (PDB: 3C6K), which causes X-linked Snyder-Robinson syndrome [28]. The substitution is predicted to decrease protein stability, and the prediction confidence is 82.32%.

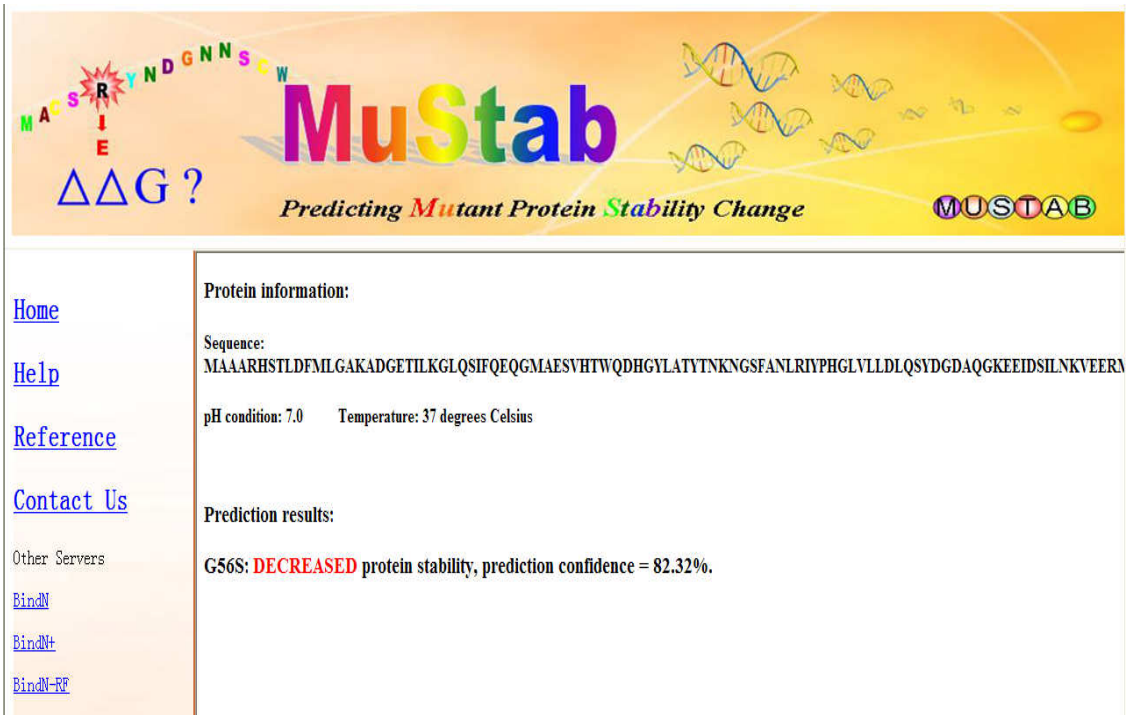

Figure 4.4 Sample output from the MuStab web server.

CONCLUSION

In this study, we have developed a machine learning method for predicting protein stability changes upon amino acid substitutions. The novelty of our method lies in the use of sequence features representing biological knowledge for input encoding. Twenty sequence features were examined for SVM classifier construction, and several of them were shown to be highly relevant for protein stability prediction. However, the SVM classifier constructed using all the twenty features did not show high predictive performance. We thus used a wrapper approach for feature selection, and identified the optimal subset of six sequence features for input encoding. The best classifier achieved

the overall accuracy of 84.59% with 70.29% sensitivity and 90.98% specificity. This SVM classifier is compared favorably in performance with the previously published models for protein stability prediction. Since the previous studies did not utilize the biological knowledge for classifier construction, our method can be used to complement the existing methods to predict the consequences of amino acid alterations in disease candidate genes and may provide useful information for elucidating the molecular mechanisms of human genetic disorders. We have thus developed the MuStab web server (http://bioinfo.ggc.org/mustab/) to make our classifier accessible to the genetics research community.

REFERENCES

1. Shirley BA, Stanssens P, Hahn U, Pace CN: Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 1992, 31(3):725-732.

2. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999, 22(3):231-238.

3. Yue P, Moult J: Identification and analysis of deleterious human SNPs. *J Mol Biol* 2006, 356(5):1263-1274.

4. Wang Z, Moult J: SNPs, protein structure, and disease. *Hum Mutat* 2001, 17(4):263-270.

5. Capriotti E, Fariselli P, Casadio R: A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* 2004, 20 Suppl 1:i63-68.

6. Capriotti E, Fariselli P, Casadio R: I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 2005, 33(Web Server issue):W306-310.

7. Cheng J, Randall A, Baldi P: Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 2006, 62(4):1125-1132.

8. Huang LT, Gromiha MM, Ho SY: iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* 2007, 23(10):1292-1293.

9. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A: ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res* 2004, 32(Database issue):D120-121.

10. Gasteiger E. HC, Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.: The Proteomics Protocols Handbook: Humana Press; 2005.

11. Kawashima S, Kanehisa M: AAindex: amino acid index database. *Nucleic Acids Res* 2000, 28(1):374.

12. White SH: Amino acid preferences of small proteins. Implications for protein stability and evolution. *J Mol Biol* 1992, 227(4):991-995.

13. Deleage G, Roux B: An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng* 1987, 1(4):289-294.

14. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH: Hydrophobicity of amino acid residues in globular proteins. *Science* 1985, 229(4716):834-838.

15. Cho MK, Kim HY, Bernado P, Fernandez CO, Blackledge M, Zweckstetter M: Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau. *J Am Chem Soc* 2007, 129(11):3032-3033.

16. Zhou H, Zhou Y: Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* 2004, 54(2):315-322.

17. Takano K, Yutani K: A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Eng* 2001, 14(8):525-528.

18. Bhaskaran R PP: Positional flexibilities of amino acid residues in globular proteins. . *Int J Pept Protein Res* 1988, 32:242-255.

19. Aboderin AA: An empirical hydrophobicity scale for alpha-amino-acids and some of its applications. *Int J Biochem* 1971, 2:537-544.

20. Jones DD: Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. *J Theor Biol* 1975, 50(1):167-183.

21. S F: Theoretical prediction of protein antigenic determinants from amino acid sequences. *Can J Chem* 1982, 60:2606-2610.

22. Zhao G, London E: An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Sci* 2006, 15(8):1987-2001.

23. Noble WS: What is a support vector machine? *Nat Biotechnol* 2006, 24(12):1565-1567.

24. Wang L, Brown SJ: Prediction of DNA-binding residues from sequence features. *J Bioinform Comput Biol* 2006, 4(6):1141-1158.

25. Wang L, Brown SJ: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006, 34(Web Server issue):W243-248.

26. Swets JA: Measuring the accuracy of diagnostic systems. *Science* 1988, 240(4857):1285-1293.

27. Bradley A: The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997, 30:1145-1159.

28. de Alencastro G, McCloskey DE, Kliemann SE, Maranduba CM, Pegg AE, Wang X, Bertola DR, Schwartz CE, Passos-Bueno MR, Sertie AL: New SMS mutation leads to a striking reduction in spermine synthase protein function and a severe form of Snyder-Robinson X-linked recessive mental retardation syndrome. *J Med Genet* 2008, 45(8):539-543.

CHAPTER FIVE

MODELING EFFECTS OF HUMAN SINGLE NUCLEOTIDE POLYMORPHISMS ON PROTEIN-PROTEIN INTERACTIONS[4]

ABSTRACT

A large set of 3D structures of 264 protein-protein complexes with known non-synonymous SNPs (nsSNPs) at the interface was built using homology-based methods. The nsSNPs were mapped on the proteins' structures and their effect on the binding energy was investigated with CHARMM force field and continuum electrostatic calculations. Two sets of nsSNPs were studied: disease annotated (OMIM) and non-annotated (non-OMIM). It was demonstrated that OMIM nsSNPs tend to destabilize the electrostatic component of the binding energy, in contrast with the effect of non-OMIM nsSNPs. In addition, it was shown that the change of the binding energy upon amino acid substitutions is not related to the conservation of the net charge, hydrophobicity or hydrogen bond network at the interface. The results indicate that, generally, the effect of nsSNPs on protein-protein interactions cannot be predicted from amino acids' physico-chemical properties using the structure-based methods alone, since in many cases a substitution of a particular residue with another amino acid having completely different polarity or hydrophobicity had little effect on the binding energy. Analysis of sequence conservation showed that nsSNP at highly conserved positions resulted in large variance of the binding energy changes. In contrast, amino acid substitutions corresponding to nsSNPs at non-conserved positions, on average, were not found to have a large effect on

[4]Teng S, Kundrotas P, Madej T, Panchenko A, Alexov E: Modeling effects of human SNPs on protein-protein interactions. Biophysics. J. 2009, 96(6):2178-2188.

binding affinity. pKa calculations were performed and showed that amino acid substitutions could change the wild type proton uptake/release and thus resulting to different pH-dependence of the binding energy.

INTRODUCTION

Each individual possesses unique characteristics reflecting their genotype, i.e. the uniqueness of the individual's DNA [1]. For example, almost all nucleotide bases (99.9%) are exactly the same in all people; however, the remaining 0.1% account for about 1.4 million individual-specific differences (single nucleotide polymorphism: SNP) that occur in humans. These differences may be within the coding or non-coding regions of DNA and may or may not result in amino acid changes, which, in turn, can either be harmless or disease causing [2]. From a computational biophysics point of view, SNPs resulting in amino acid changes (non-synonymous SNP: nsSNP) are of particular interest because such changes should affect the stability of proteins and protein-protein complexes.

From a biological perspective, the major factor contributing to the complexity of biological systems is the high degree of connectivity on the molecular scale. In particular, many proteins responsible for cellular functions rely on interactions with other proteins to perform these functions. If the structures of the corresponding protein-protein complexes are available, then we will have the opportunity to apply theoretical biophysical methods to model the energetics of protein-protein complexes [3-9] and apply the results in structure-based drug design [10]. Thus, understanding protein-protein interactions and

their roles in cell function will help reveal the molecular mechanisms of protein recognition and model of the effect of perturbations on biological network, in particular, the effects of nsSNPs on protein-protein interactions [11-14].

The effects caused by nsSNPs can be broadly grouped into four distinctive categories [15] (although the effects may be mutually dependent) depending on what type of system or process have been affected by nsSNPs: (a) protein folding, stability, flexibility and aggregation; (b) functional sites, reaction kinetics and dependence on the environmental parameters, such as pH, salt concentration and temperature; (c) protein expression and subcellular localization and; (d) protein-small molecule, protein-protein, protein-DNA and protein-membrane interactions (see review and references within [15]). Among these categories, the effect of nsSNPs on protein stability [16-18] attracted most of the attention of scientific community. The mechanisms of the effect of nsSNPs on protein stability could vary from geometrical constraints (the mutation of a small side chain to a bulky side chain in the protein interior), to physico-chemical effects (replacement of hydrophobic residue with polar residue), to the reversal of a charge within a salt bridge, or to the disruption of hydrogen bonds [19]. For example, the nsSNPs resulting in changes of functionally important residues should be almost always deleterious as they would block protein function [20] [21]. However, since there are only a few functional residues within an entire protein sequence, the probability for such mutations is low [22]. The possibility of a nsSNP affecting the subcellular location of a corresponding protein was reported in a recent study which showed that in about one percent of the cases the disease is caused by protein subcellular delocalization [23]. In

addition to the above mentioned effects, nsSNPs can change the kinetics of the corresponding reactions as it was experimentally shown in case of patients with chronic lymphocytic leukemia [24], inflammatory diseases [25] or to affect the pharmacokinetics [26], however modeling these effects is computationally difficult. However, although the studies of consequences of nsSNPs on proteins have drawn much attention recently, the effect of nsSNPs on protein-protein interactions has not been extensively investigated. Perhaps this is due to the lack of sufficiently many 3D structures of protein-protein complexes for which nsSNPs are known.

The progress recently made in experimental 3D structure determination, led by the Structural Genomic Initiatives [27], in addition to advances in computational modeling [28, 29] made it possible to predict the effects of nsSNPs by mapping them on the corresponding structures or on the protein and protein-protein models. Indeed, structural information was used in many studies to reveal the role of SNPs on protein function and stability. A recent study on human nsSNPs and disease-associated mutations in orthologous genes revealed that approximately 70% of disease-associated mutations were in protein sites that most likely affect protein function [30-33]. Moreover, it was found that disease mutations are much more likely to occur at sites with low solvent accessibility [32]. Recently, a structure-based approach that models residue-residue interaction networks was reported [34]. It applied graph theoretical measures to predict the residues that are important for structural stability. These results imply that nsSNPs impact protein function and stability by affecting their structures, which in turn might cause changes in protein-protein or protein-ligand interactions.

It should be mentioned that most of the efforts in the field so far have been aimed at predicting deleterious mutations, since such predictions could be used for early diagnostics and potential drug discovery [23, 31, 32, 35-38]. However, the goals of our study are: (a) to investigate the possibility that disease-causing and harmless nsSNPs affect protein-protein interactions differently, and (b) to reveal the basic principles of the effects of naturally occurring interfacial nsSNPs on protein-protein interactions. The rationale behind our approach is that any mutation at a protein-protein complex interface should, in principle, affect somehow the binding energy and even harmless nsSNPs can also cause dramatic changes in the phenotype resulting in natural differences among individuals. To deduce the effect of nsSNPs on protein function, further investigation of the effect of nsSNPs on protein-protein interaction network is needed, combined with detailed analysis of the importance of the perturbed interactions for normal cellular function.

In this study, we use homology modeling to construct 3D models of a large number of protein-protein complexes (264) with known nsSNPs at their interfaces. The effect of amino acid substitution resulted from nsSNPs on the protein-protein binding energy was calculated using a standard force field (CHARMM [39]), in contrast to previous studies that applied descriptors or semi-empirical functions. In addition, specific attention was paid to possible ionization changes and charge reorganization caused by the nsSNP mutations. The calculated effects are grouped into categories that describe several distinctive mechanisms of nsSNPs affecting the energetics of protein-protein interactions. The role of charge relaxation is also investigated.

METHODS

Sequence alignment, template detection and model building

The first task was to extract query amino acid sequences associated with nsSNPs and to search for available 3D structures or for 3D structures that are homologous to the query sequences. The locus-id files for humans were downloaded from build 126 of the dbSNP database, which contains the SNPs associated with gene names and locations on genes. These files also included accessions for protein sequences associated with the SNPs. The protein sequences, which were found to be associated with SNPs, were compared against the set of human protein structures (potential structural templates) (NCBI MMDB) [40], using Blast algorithm [41]. Those human structures which were found at an E-value of 10e-5 or better were kept resulting in 5.6 millions alignments. If a 3D structure of a query protein was available, no modeling was required. Query proteins that matched any of the entries in the OMIM database [42-44] were marked as "annotated" disease-causing. The rest of the entries were considered undetermined with respect to possible disease association and are referred to in the manuscript as "non-annotated" or "non-OMIM".

At the second stage of processing, additional criteria were used requiring that 80% of the query sequence to be mutually aligned with the structural template (nsSNPs that were not mapped in the alignment were discarded). Only templates corresponding to protein-protein (or domain-domain) complexes were used for modeling 3D structures of nsSNP containing sequences. During this procedure, we recorded whether or not the SNP was on the interface for each chain/domain pair. It was done using query-template Blast

alignments. Interface residues were defined as those being 8Å from each other (distance was measured between C-alpha atoms) on different chains/domains[45]. These positions were flagged as interfacial residues.

The detected templates and corresponding sequence alignments were used as input for the homology modeling. The 3D models were built with program NEST using the sequence alignment between queries and structural templates [46]. Identical alignments were discarded. The number of models built for different degrees of modeling difficulty were as follows: (1) 1257 models were built by side chain replacement where query and template sequences differed only by a few residues and the models were built by mutating corresponding residues in the original chain and (2) 5274 models were built with the NEST program. Because of the restrictive alignment criteria applied above, in most of the cases, the alignment had very few gaps/insertions, and thus the models were very close to the template structures. In total, 6531 protein models were constructed which corresponded to the first allele (the first allele in case of OMIM is the dominant allele, while in case of non-OMIM it is simply the first allele in the list). Then the monomeric proteins models were joined to the corresponding partners using the 3D structure of template protein-protein complex. The models of complexes were then evaluated according to the flagged interfacial positions, and only models with nsSNPs occurring at the interface of protein-protein complexes were retained for our study, resulting in 264 model structures.

Energy minimization

The structures of the 264 complexes were subjected to the TINKER package [47] using the CHARMM27 force field parameters [39]. The minimization was done running the TINKER's *minimize.x* module. The *minimize.x* module performs energy minimization using the Limited Memory BFGS Quasi-Newton Optimization algorithm [47]. The implicit solvent was modeled using the Still Generalized Born model [48], and the internal dielectric constant was set to 1.0 to be consistent with the CHARMM27 force field parameters [49]. The convergence criteria applied was RMS gradient per atom = 0.01. For energy minimization calculations, we utilized a High Throughput Distributed Computing Resource, CONDOR, originally developed at the University of Wisconsin-Madison (www.cs.wisc.edu/condor), which is now available at Clemson University with more than 1,080 single CPUs of computational power.

The minimized 3D structures of the complexes with amino acids corresponding to the first reported allele in the dbSNP database were then used to generate the corresponding nsSNP mutations. Utilizing the SCAP program [50], the mutations, corresponding to either the second allele in the dbSNP database or the disease-causing nsSNP in OMIM database, were introduced using the above minimized model 3D structures, while keeping the rest of the structure rigid, including the hydrogen atoms. In case of homooligomeric-complexes, the nsSNP mutations were introduced on both monomers. Then, the resulting 3D structures were minimized again with TINKER using the same protocol that was described above.

Binding energy calculations

The binding energy was calculated with the so-called rigid body approach keeping the structures of the monomers as they were in the complexes. Such an approach is advantageous because the internal mechanical energies of the unbound and bound monomers are the same and do not have to be included in the calculations of the binding energy. Thus, the single point calculations result in binding energy:

$$\Delta\Delta G(binding) = \Delta G(complex) - \Delta G(A) - \Delta G(B) \qquad (5.1)$$

where *ΔG(complex), ΔG(A) and ΔG(B)* are the unfolding free energy for the complex, monomer A and monomer B, respectively. The total binding energy and its two components (electrostatics and van del Waals) were analysed. The electrostatic component of the binding energy is the sum of the Coulombic and reaction field energies as described in detail in [51, 52]:

$$\Delta G_{el}(X) = G(coul) + \Delta G(rxn) \qquad (5.2)$$

where *X* stands for the complex, *A* and *B* monomers, respectively. *G(Coul)* is the Coulombic interaction energy, and *G(rxn)* is the reaction field energy, which is calculated with Delphi program [51, 52].

The total binding energy is:

$$\Delta G_{tot}(X) = \Delta G(bonds) + \Delta G(vdW) + \Delta G(el) \qquad (5.3)$$

where *ΔG(bonds)* are the bonded energy terms, *ΔG(vdW)* is the van der Waals energy and *ΔG(el)* is the Coulombic interactions and solvation energy calculated with the Generalized Born (GB) model. However, since we adopted the rigid body approach,

$\Delta G(bonds)$ for the complexes and free monomers is the same and cancels in eq. (5.3). All of the above energy terms were calculated with the *analyze.x* module in TINKER. The non-polar component of the binding energy was not included in the calculations because the single point mutation is not expected to change the binding interface significantly.

Changes in protein stability caused by the nsSNP mutation were calculated with respect to the energy of the target (the first reported allele or wild type allele in case of OMIM nsSNPs) protein. The corresponding quantity is $\Delta\Delta\Delta G(snSNP)$, as described below:

$$\Delta\Delta\Delta G(snSNP) = \Delta\Delta G(target:binding) - \Delta\Delta G(snSNP:binding) \qquad (5.4)$$

The changes of the total binding energy $(\Delta\Delta\Delta G_{tot}(nsSNP))$, as well as the change of its vdW $(\Delta\Delta\Delta G_{vdw}(nsSNP))$ and electrostatic $(\Delta\Delta\Delta G_{el}(nsSNP))$ components are analyzed in this work. If the change is negative, this indicates that the nsSNP mutation weakens the affinity and destabilizes the complex, while if the change is positive then the mutant binding is tighter.

Multiple sequence alignment

Protein sequences from different species were downloaded from the NCBI Entrez database, using GENE search option and submitting each of the gene's ID as a query. Only cases for which a protein was found in more than four species were considered, and the MSAs were built resulting in 227 out of the total 264 sequences. We used EBI's ClustalW2 web service (http://www.ebi.ac.uk/Tools/clustalw2/index.html) to perform multiple sequence alignments (MSA).

pKa calculations of the ionizable states and proton uptake/release

The pKa values of the ionizable groups were calculated using the Multi Conformation Continuum Electrostatics (MCCE) method as previously described [53-55]. Recently, we demonstrated that MCCE can be utilized to calculate pKa's using 3D structures that were built by homology [56]. Calculations were performed for all 264 protein complexes corresponding to the first allele, and another set of pKa calculations were done for the protein complexes with corresponding nsSNP mutation. The calculations were also performed on the corresponding unbound monomers, which structures were taken from the corresponding protein-protein complex. These results were used to predict the changes of the titratable groups' ionization states caused by complex formation. For each complex, we calculated the difference of the net charge *(Δq(X))* of the complex and of the unbound monomers, called proton uptake/release:

$$\Delta q(X) = q(X : complex) - q(X : A) - q(X : B) \qquad (5.5)$$

where X is the first allele or nsSNP variant, and q is the net charge of the complex and of monomer A and B, respectively, calculated with MCCE at a pH of 7.0. We chose a pH of 7.0 because there was no information of what is the physiological pH for each of the proteins studied in this manuscript. In addition, we analyzed the proton uptake/release difference between complexes with the first allele and the nsSNP variant:

$$\Delta\Delta q = abs(\Delta q(\text{dominant allele}) - \Delta q(\text{nsSNP})) \qquad (5.6)$$

P-value calculations

The P-values were calculated performing t-test [57-59]. The distributions of the corresponding changes of the binding energy and its components in case of OMIM and non-OMIM sets were checked against the null hypothesis. Large P-value indicates that the corresponding distribution is similar to the normal distribution (null hypothesis), while small P-value points out a deviation from random distribution. Typical cut-off for P-value is 0.01, i.e. distribution with P-value smaller than 0.01 is considered significantly different from random. The distribution of the variance of $\Delta\Delta\Delta G_{tot}(nsSNP)$ and $\Delta\Delta\Delta G_{el}(nsSNP)$ was checked against the null hypothesis that assumes equal variances. The SI% scale was divided into five bins, corresponding to cases with SI% smaller than 20%, 20% < SI% <40%, 40% < SI% < 60%, 60% < SI% < 80, and 80% < SI% < 100%. The variance of the corresponding energies was calculated within each of the bins and the resulting P-value evaluated. In case of $\Delta\Delta q$, six bins were considered: $0.00 < \Delta\Delta q < 0.05$, $0.05 < \Delta\Delta q < 0.10$, $0.10 < \Delta\Delta q < 0.15$, $0.15 < \Delta\Delta q < 0.20$, $0.20 < \Delta\Delta q < 0.25$ and $\Delta\Delta q > 0.25$. Then, the variance of the corresponding energies within these bins and the P-value were calculated.

RESULTS AND DISCUSSION

Distribution of binding energy

The changes in the total binding energy and its electrostatic and vdW components due to the nsSNPs were calculated for all complexes in the dataset (Figure 5.1, Table 5.1). The distributions of $\Delta\Delta\Delta G_{tot}(snSNP)$ for OMIM and non-OMIM cases are shown in

Figure 5.1a. It can be seen that the distributions have similar shapes, showing slight tendency toward negative values. The mean values of electrostatic ($\Delta\Delta\Delta G_{el}(snSNP)$) and vdW ($\Delta\Delta\Delta G_{vdw}(snSNP)$) components of the binding energy changes are statistically different for OMIM and non-OMIM cases (P-values are less than 0.006 and 0.01 respectively), although this is not the case for the total binding energy. Figure 5.1b shows the distribution of $\Delta\Delta\Delta G_{el}(snSNP)$ for both OMIM and non-OMIM cases. One can see the long negative tail of the distribution of OMIM cases for which nsSNP substitutions destabilize binding. Moreover, the mean of OMIM distribution of electrostatic energy is significantly different from zero and shifted towards negative values although this is not the case for non-OMIM distribution of electrostatic component (Table 5.1). This indicates that, overall, there is a tendency for OMIM nsSNP substitutions to weaken electrostatic component of the binding energy, although there are many examples where disease nsSNPs make binding tighter as well. The effect is less pronounced for the total binding energy.

Table 5.1: Parameters of distributions of total binding energy difference and its components together with the corresponding P-values (the null hypothesis that mean value is greater or equal to zero is rejected if P-value is less than 0.01)

Group	No	$\Delta\Delta\Delta G_{tot}$			$\Delta\Delta\Delta G_{vdw}$			$\Delta\Delta\Delta G_{el}$		
		mean	std	P-value	mean	std	P-value	mean	std	P-value
OMIM	45	-1.65	3.80	0.003	-1.03	3.32	0.02	-2.35	5.51	0.003
Non-OMIM	219	-0.70	4.36	0.009	0.14	3.03	0.75	-0.45	4.39	0.06
Polar (P)	62	-0.27	3.77	0.28	0.38	3.94	0.77	-0.83	4.74	0.09
Charge (C)	76	-2.01	6.38	0.004	-0.33	2.25	0.1	-1.37	6.59	0.04
Small (S)	94	-0.74	2.39	0.002	-0.03	2.49	0.45	-0.78	2.58	0.002
Hydrophobic (H)	32	0.32	2.50	0.77	-0.36	4.46	0.32	0.74	3.23	0.09

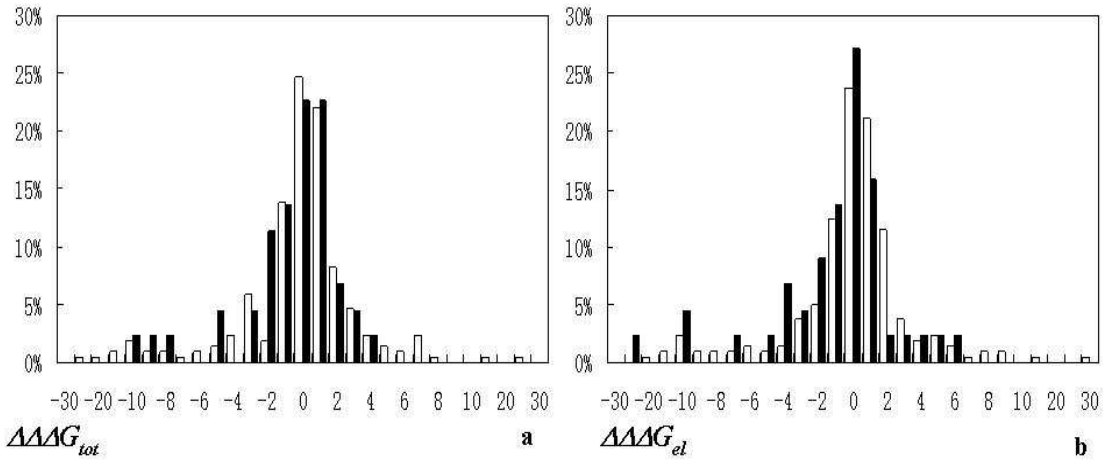

Figure 5.1: ROC Distribution of $\Delta\Delta\Delta G_{el}$*(nsSNP)* and $\Delta\Delta\Delta G_{tot}$*(nsSNP*) for OMIM and non-OMIM cases. OMIM: black bars, non-OMIM: white bars.

From an electrostatic point of view, replacing the wild type amino acid (dominant allele) at a protein-protein interface with another amino acid (amino acid which corresponds to nsSNP) is expected to be a destabilizing event. Indeed, in our previous study of 654 protein-protein and domain-domain complexes, we demonstrated that the electrostatic component of the binding energy tend to be optimized [60] with respect to random shuffling of the amino acid sequences of the corresponding binding partners. Thus, since wild type (dominant allele) interactions across the interface are optimized, any change should make the binding affinity weaker. Indeed, the destabilization effect upon disease substitutions is the most pronounced in case of electrostatic component of binding energy ($\Delta\Delta\Delta G_{el}$ distributions is shifted toward negative values with P-value of less than 0.003). However, the tendency of OMIM mutations to destabilize electrostatic component of the binding energy is not very strong which perhaps stems from the fact that nsSNP substitutions are not random, rather they are constrained mutations accepted

by the cell. At the same time, for non-OMIM substitutions the electrostatic component should be optimized for both alleles and consequently the mean of $\Delta\Delta\Delta G_{el}(nsSNP)$ is not statistically significant different from zero (P-value is 0.06).

Despite the differences, in majority of the cases, both OMIM and non-OMIM substitutions were calculated to have little effect on the binding. Since we investigate nsSNP substitutions at the interface of protein complexes, such an observation deserves further investigation. The next sections investigate possible patterns and correlations between different types of amino acid substitutions and their calculated effects on the binding energy.

Effect of nsSNPs on binding energy with respect to amino acid characteristics

In this section, four different classes of amino acids were considered based on the amino acids' physico-chemical properties: polar (S, T, H, N, Q, Y), charged (E, D, K, R), hydrophobic (W, I, L, M, F) and small (P, A, G, C, V). We adopt such a simplified classification to ensure that each class has enough representatives in our dataset. Of course, many other classifications exist, including more detailed definitions of the subgroups. Below we investigate the effects of nsSNP mutations on the $\Delta\Delta\Delta G_{tot}(snSNP)$, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ and $\Delta\Delta\Delta G_{el}(nsSNP)$ separately for each class (more detailed analysis including analysis of the effects of substitutions between classes is given in the supplementary results).

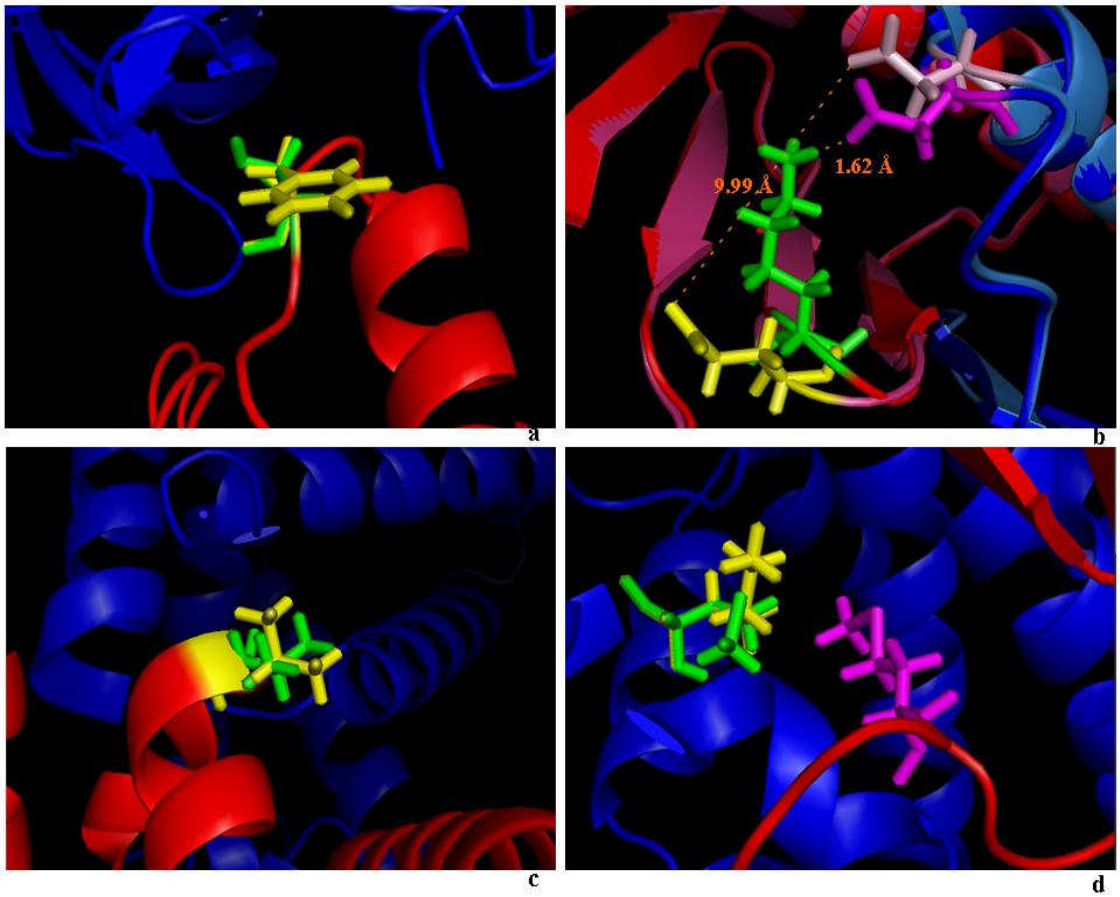

Figure 5.2: Illustration of nsSNPs at interface of protein-protein complexes. (a) TTR (transthyretin, Gene ID: 4507725) Red: A Chain, Blue: E Chain. Green: Ser in A85, Yellow: F in A85, Magenta: N in E63. (b) DYNLRB1 (Roadblock-1, Gene ID: 7661822) Red: A Chain of target, Light Red: A Chain of SNP variants. Blue: B Chain of target, Sky Blue: B Chain of SNP variant, Green: K in A75, Yellow: E in A75, Magenta: D in B61 of target, Pink: D in B61 of SNP variant. (c) HBB (beta globin, Gene ID: 4504349) Red: B Chain, Blue: C Chain. Green: V in B34, Yellow: L in B34. (d) GSTM2 (glutathione S-transferase M2, Gene ID: 4504175) Red: A Chain, Blue: B Chain. Green: M in A130, Yellow: K in A130, Magenta: M in B50

Binding energy changes caused by a substitution of a ***polar*** amino acid

There are 62 cases in our dataset for which a polar residue corresponding to the first allele and located at the interface of the protein-protein complex is substituted by other variant (Table 5.1). Overall, there is no statistically significant bias for energy to be shifted upon substitution towards lower or higher values.

From an electrostatic point of view, a *polar→another* amino acid substitution tends to be an unfavorable event in the majority of cases (P-value=0.09). In another words, removal of a polar group at the interface, despite structural refinement, makes electrostatic binding energy less favorable. Further analysis of such cases showed that a removal of a polar residue disturbs the hydrogen bond network at the interface. Substitution of a polar residue with either a small, charged or hydrophobic groups tends to make electrostatic component of binding weaker. A small residue will create energetically unfavorable cavities, a charged residue will pay large desolvation penalty and a hydrophobic residue will not be able to provide the required hydrogen bonds. However, exceptions are cases when a polar group is replaced by another polar residue whose side chain can satisfy the required geometry. In the last case, the electrostatics may not change or even become more favorable.

A particular example of *polar→hydrophobic* substitution is shown in Figure. 2a. It demonstrates that removal of a polar residue and substitution with a hydrophobic residue results in the placement of the hydrophobic side chain in a polar environment, an event that weakens the binding affinity. A typical case is Transthyretin (TTR), which is a plasma protein that binds retinol and thyroxine. Many distinct forms of amyloidosis are

related to different nsSNPs in TTR. For example, the nsSNP (refSNP ID: rs11541784) results in a change of the polar (Ser) residue into a hydrophobic residue (Phe). The nsSNP Phe residue is located in a polar environment and reduces the binding affinity by 0.717 kcal/mol.

Binding energy changes caused by a substitution of a ***charged*** amino acid

There are 76 cases in our dataset in which a charged residue located at the interface of the target protein-protein complex is substituted in the nsSNP variant (Table 1). The values of the means of $\Delta\Delta\Delta G_{tot}(nsSNP)$ and its electrostatic component $\Delta\Delta\Delta G_{el}(nsSNP)$ are negative and this bias is statistically significant (P-values 0.004 and 0.04, respectively) which means that the target protein-protein complexes are more stable compared to the nsSNP variants.

Substituting a charged with another residue is, overall, an unfavorable event with respect to protein-protein association (Table 5.1). Removal of a charged residue that forms a salt bridge across the interface in the target complex leaves the charged partner without favorable pair-wise interactions. The remaining charged residue pays a huge desolvation penalty upon complex formation, which in the nsSNP variant may not be compensated by favorable pair-wise interactions. This provides an intuitive explanation why distributions of both the $\Delta\Delta\Delta G_{tot}(nsSNP)$ and $\Delta\Delta\Delta G_{el}(nsSNP)$ are shifted toward negative values .

The mutation of a charged amino acid to another charged amino acid (*charged*→*charged*) is an interesting case. The mutation could preserve the charge (Asp

⇔ Glu; Lys ⇔ Arg) or invert the charge (Asp,Glu ⇔ Lys, Arg). Presumably, a mutation that preserves the charge should have a lesser effect on the binding energy as compared with charge-reversal mutations. However, our analysis showed that this is not always the case. Overall, all mutations of the target charged residue to another charged residue were found to be unfavorable events (Table 5.1). Even in the case of Glu to Asp substitutions, like aldolase B (s2525, Glu to Asp in position 64), which is a mutation (refSNP ID: 2854709) that preserves the net charge of the complex, the change of the binding energy is huge: $\Delta\Delta\Delta G_{tot}(nsSNP)$ = -9.06 kcal/mol, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ = -1.58 kcal/mol and $\Delta\Delta\Delta G_{el}(nsSNP)$ = -11.30 kcal/mol. This is due to the fact that the side chain of Asp is shorter than the Glu side chain, and the nsSNP introduced Asp cannot form a strong salt bridge with the original partner Lys in position 270 of the other chain in this homo-dimer complex. Another example (Figure 5.2b) is the case of charge reversal in Roadblock-1 (DYNLRB1), which is a homo-dimeric protein that may be involved in tumor progression, as the up-regulation of this gene is associated with hepatocellular carcinomas. The corresponding nsSNP (refSNP ID: rs11537531) of this protein results in the change of a Lys amino acid to a Glu amino acid at the complex's interface. In the target protein complex, the distance between Lys75 from chain A and its partner Asp61 from chain D is only 1.62 Å, resulting in a very strong hydrogen bond and pair-wise electrostatic interactions. However, in the nsSNP variant, the positively charged Lys is replaced by Glu, a negatively charged residue. Due to minimization, the distance between the nsSNP residue and the original Asp61 from chain D increases to 9.99Å due to the repulsive charge-charge interaction between the two negatively charged groups (Figure

5.2b). This reduces the effect, but the binding energy is still much less favorable as compared with the dominant allele. The corresponding energy changes are $\Delta\Delta\Delta G_{tot}(nsSNP)$ = -11.13 kcal/mol, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ = -4.42 kcal/mol and $\Delta\Delta\Delta G_{el}(nsSNP)$ = -3.08 kcal/mol. This is an example of a structural relaxation that reduces the effects of charge reversal.

Binding energy changes caused by a substitution of a ***small*** amino acid

There are 94 cases in our dataset for which a small residue located at the interface of the target protein-protein complex is substituted into the nsSNP variant (Table 1). Overall, the total binding energy and electrostatic components are statistically significant (both P-values are 0.002) shifted toward negative values which indicates that nsSNP destabilizes the complex.

Substitution of a small with another amino acid almost always will result in sterical clashes. The volume of a small amino acid is much smaller that the volume of the other residues. Thus, there will be no room for bulky amino acid side chain at the interface. Such a replacement will cause distortion of the interface and will weaken the binding (Table 5.1). A typical example is the histidine triad nucleotide binding protein 1 (HINT1), Gene ID: 4885413. The nsSNP codes for Gly to Arg substitution in position 92 of B chain. The substitution introduces a new charged residue, which pays large desolvation penalty and the resulting change in the electrostatic component of the binding energy $\Delta\Delta\Delta G_{el}(nsSNP)$ is -9.23 kcal/mol).

However, there are also opposite examples, indicating that protein complexes can tolerate small amino acid substitutions at the interfaces. A typical example is Human β-globin (HBB), which regulates developmental expression. The corresponding nsSNP (refSNP ID: rs1141387) in this protein replaces a Val residue with a Leu amino acid. Despite the difference in these two amino acids' volumes, the structure of the complex does not change by much, resulting in very small energy differences: $\Delta\Delta\Delta G_{tot}(nsSNP)$ = -0.16 kcal/mol $\Delta\Delta\Delta G_{vdw}(nsSNP)$ = -0.44 kcal/mol and $\Delta\Delta\Delta G_{el}(nsSNP)$ = -0.00 kcal/mol (Figure 5.2c). The main reason for this is that both side chains are partially exposed to the solution, and there is room for a larger Leu side chain.

Binding energy changes caused by a substitution of a ***hydrophobic*** amino acid

There are 32 cases in our dataset in which a hydrophobic residue located at the interface of the target protein-protein complex is substituted by the nsSNP variant (Table 1). The mean values of all energy distributions are not significantly different from zero. In general, substituting a hydrophobic residue at the interface with another residue does not have large effect on protein-protein binding. Perhaps this is due to the fact that hydrophobic groups do not form specific interactions. Thus, the effect of a replacement of a particular hydrophobic side chain with another residue depends on geometry of the interface and the ability of the substituted side chain to form new interactions. For example, a polar or charged residue, substituting a hydrophobic one, could increase the binding affinity only if the corresponding residue manages to create new favorable interactions across the interface. If this does not occur, then the mutation should weaken

the binding. Such a case is shown in Figure. 2d. Glutathione S-transferase M2 (GSTM2) is an important enzyme that contributes to the metabolism of phase II biotransformation of xenobiotics. The corresponding nsSNP (refSNP ID: rs1056799) changes the target amino acid Met to Lys in position A130. However, the new charged residue cannot form favorable interactions with any other residue across the interface since it is in a hydrophobic environment. As a result, the solvation loss cannot be compensated for, and the mutation weakens the binding.

Correlation of the calculated effect on the binding affinity and residue conservation

Multiple sequence alignments (MSA) were used for phylogenetic analysis and for determining the evolutionary relationships between different species. Only positions corresponding to interfacial sites were considered. A position in the MSA that is totally or highly conserved indicates strong evolutionary constraints, and the substitution of such a highly-conserved amino acid is expected to have significant effects on protein structure, function and interactions. In contrast, an amino acid that is not conserved among different species is, perhaps, not crucial for the structure, function and interactions of that particular protein complex.

We began our analysis with a case corresponding to a highly conserved site. Position B34 in Human β-globin (HBB) is totally conserved among the species (Figure 5.3a). The nsSNP causes a mutation that changes Val residue to Leu. As result, the total binding energy, van der Waals and electrostatic components are more favorable in the target complex compared with the nsSNP variant. The corresponding changes of the

binding energy are $\Delta\Delta\Delta G_{tot}(nsSNP)$ = -0.94 kcal/mol, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ = -0.10 kcal/mol and $\Delta\Delta\Delta G_{el}(nsSNP)$ = -1.12 kcal/mol.

Another example is GSTM2, glutathione S-transferase M2 (Figure 5.3b). Position A130 is not conserved; in humans it is a Met residue, however, in other species the same position is a Lys amino acid. The nsSNP induces a Met → Lys change in the human protein, a mutation that is already seen in other species. Perhaps this explains why such a drastic change (a hydrophobic to a charged group) has little effect on the binding affinity. The corresponding changes of the binding energy are $\Delta\Delta\Delta G_{tot}(nsSNP)$ = -0.53 kcal/mol, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ = 0.28 kcal/mol and $\Delta\Delta\Delta G_{el}(nsSNP)$ = -0.26 kcal/mol.

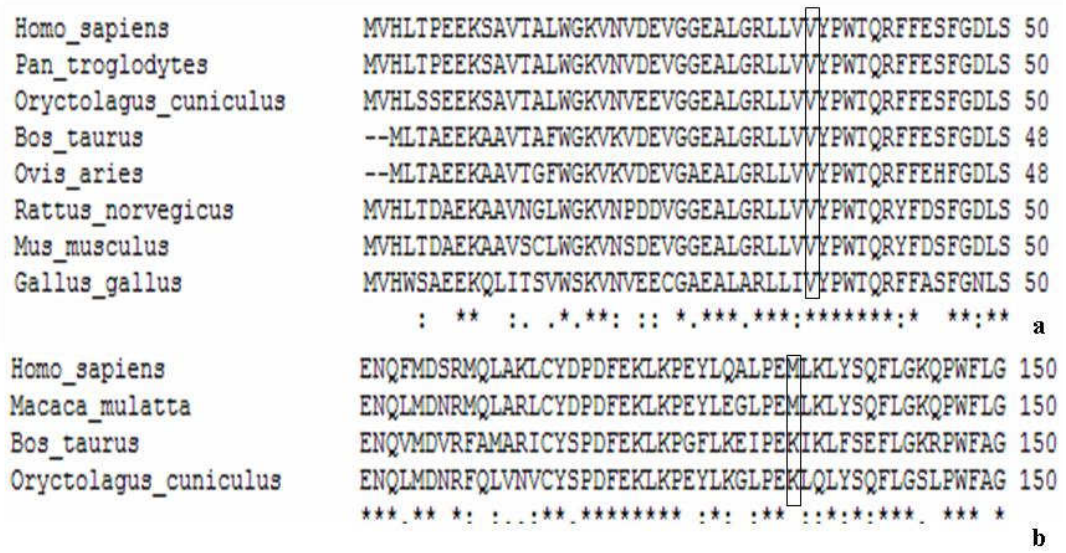

Figure 5.3: Multiple sequence alignment (MSA). The blank frame is nsSNP position. (a) HBB (beta globin, Gene ID: 4504349). (b) GSTM2 (glutathione S-transferase M2, Gene ID: 4504175)

The magnitude of the binding energy change as a function of the degree of conservation is shown in Figure 5.4. It can be seen that as the degree of conservation increases (calculated in terms of percent identity, SI%) the maximal amplitude of both the $\Delta\Delta\Delta G_{tot}(nsSNP)$ and the $\Delta\Delta\Delta G_{el}(nsSNP)$ increases as well (illustrated by the broken lines in Figure 5.4). The effect culminates at high SI% (SI% > 80%) where the variance of the magnitude of both the $\Delta\Delta\Delta G_{tot}(nsSNP)$ and the $\Delta\Delta\Delta G_{el}(nsSNP)$ is significantly different, i.e. the null hypothesis about the equality of variances between the bins was rejected with P-value<0.00001 (see Methods section). Note that this corresponds to significant variance of the binding constant resulting to either increase/decrease or no change of the affinity. The points located close to the horizontal axis and corresponding to highly conserved positions (Figure 5.4) indicate that in some cases, a mutation of a highly conserved amino acid may not affect the binding affinity. In these cases, the effect depends on the geometry of the interface and where the site is situated. These highly conserved sites are predominantly located at the periphery of the binding interface and apparently are not important for the binding affinity. Figure 5.4 provides indirect support demonstrating that the calculated effects are reasonable, since no large binding energy change was calculated to be associated with nonconserved positions in the MSA.

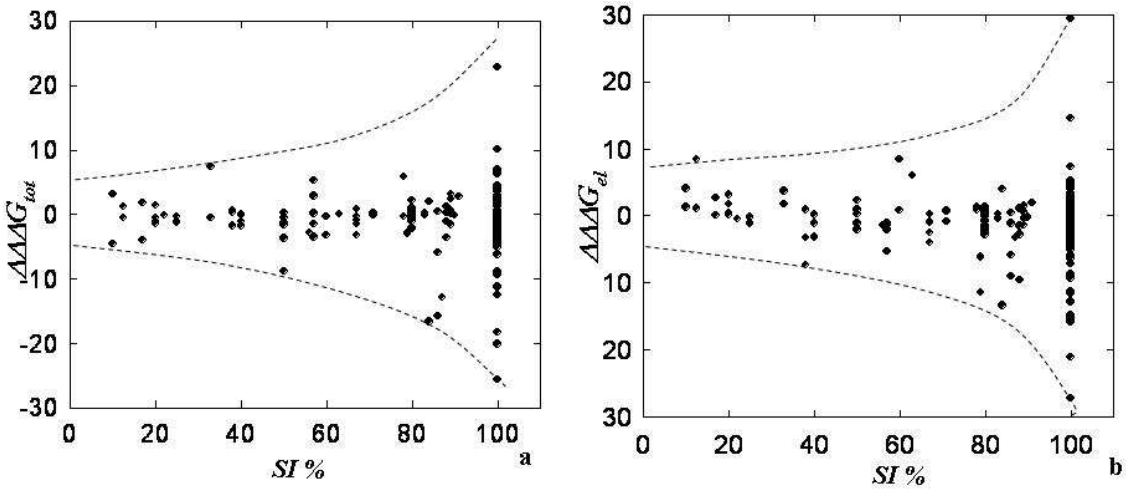

Figure. 5.4: The change of the binding energy as a function of the amino acid conservation (SI%). The broken lines are guides for the eye and follow the maximal amplitude of binding energy change). (a) $\Delta\Delta\Delta G_{tot}(nsSNP)$. (b) $\Delta\Delta\Delta G_{el}(nsSNP)$

Effect of nsSNPs on proton uptake/release

Figure 5.5 shows the change of the corresponding binding energy as a function of the absolute difference of the proton uptake/release for target complexes and an nsSNP variant calculated at pH = 7.0. No correlation between either the magnitude or variance of the binding energy change and ΔΔq was found. At the same time, it can be seen that most $\Delta\Delta q$ are close to zero, indicating that at least around a pH of 7.0 the pH-dependences of the binding energy are the same for the target complex and the nsSNP variant. However, this is not necessarily the case for the entire pH-dependence. At the same time, there is significant percentage of cases in which the $\Delta\Delta q$ is different from zero. This indicates that nsSNP mutations not only change the binding energy but also result in a different

pH-dependence of the binding. This could have a significant physiological importance; however, there is practically no experimental data available for comparison.

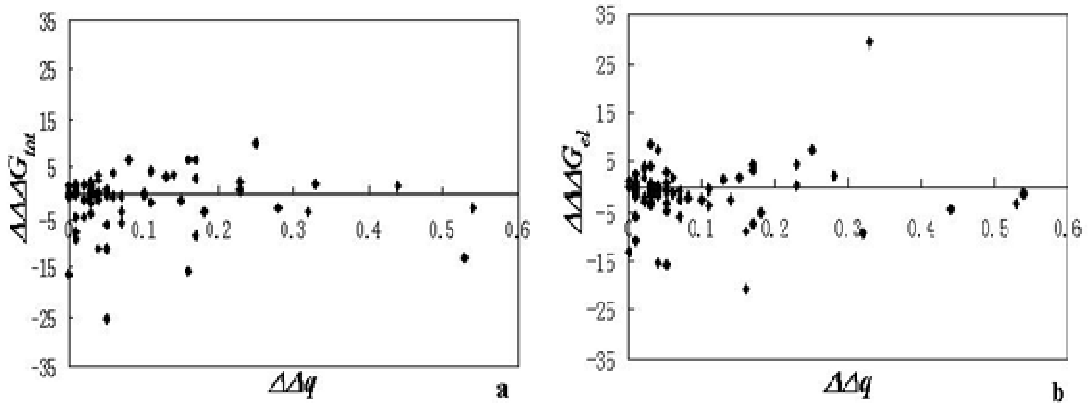

Figure 5.5: The change of the binding energy as a function of calculated proton uptake/release (absolute value of ΔΔq). (a) $\Delta\Delta\Delta G_{tot}$*(nsSNP*). (b) $\Delta\Delta\Delta G_{el}$*(nsSNP)*

In general any substitution can lead to ionization changes. The above results indicate that an amino acid substitutions corresponding to nsSNPs not only change the binding energy but could also result in changes in the ionization states of the titratable groups. Such an effect could occur not only when a titratable group is involved in the target→ nsSNP mutation but could also occur in each of the other cases as well. This is because any substitution changes the geometry of the interface and thus affects the electrostatic potential of all ionizable residues. However, in this study we did not perform charge relaxation, i.e., no attempt was made to adjust the residues' ionization states according to the pKa calculations because the calculated proton uptake/release is a fractional number. Modeling fractional ionization in single point calculations is impossible and any attempt

would be an error (see for details [61]). However, a more sophisticated approach involving ensemble presentation could take into account these ionization changes and will result in a reduction of the magnitude of the energy change caused by the nsSNP mutation. Thus, all of the data points (Figure. 5.5) corresponding to $\Delta\Delta q$ that are significantly different from zero may get closer to $\Delta\Delta\Delta G\ (nsSNP) = 0$, i.e., closer to the horizontal axis. Perhaps this is an effect that occurs *in vivo* and results in toleration of nsSNP mutations. Site-directed mutagenesis experiments and complementary numerical calculations have proven the charge-compensatory effect [62-64]. Perhaps, the charge-compensatory is the reason that maximal $\Delta\Delta q$ (Figure. 5.5) is only about 0.6 units, despite that some nsSNPs cause charge reversal.

CONCLUSION

This analysis is focused on nsSNPs located at protein-protein interfaces. Protein-protein interactions are essential for cell function, and nsSNPs affecting these interactions are expected to have significant impacts on the protein interaction network. Indeed, our analysis showed that OMIM and some non-OMIM nsSNP might have a significant effect on binding energy especially on the electrostatic component. Although the effect is statistically significant, the majority of amino acid substitution corresponding to nsSNP does not affect the binding affinity by much. This observation should be taken with caution. A small change of the binding affinity by a kcal/mol or even less could still disrupt the functionality of the interaction network or change the kinetics of the

corresponding reaction [24, 25]. However, investigating this effect requires modeling protein-protein networks, a task that is far beyond the goals of the present study.

Two data sets were considered in this study: nsSNPs that are known to be disease-causing (OMIM dataset) and nsSNPs that were not annotated to be disease-causing (non-OMIM). The distributions of the change in the binding energy and its components in both the OMIM and non-OMIM cases were found to be different although the difference is small. However, looking at the electrostatic component of the free energy we found that it is significantly shifted toward negative values for OMIM nsSNP, this is not the case for non-OMIM nsSNPs. This indicates that disease-causing nsSNPs tend to destabilize electrostatic component of protein binding energy, in contrast with non-OMIM nsSNPs.

Although large number of nsSNPs did not affect protein interactions by much (perhaps this shows the plasticity of protein interfaces and their ability to tolerate amino acid changes), an even larger fraction of the nsSNPs did affect the affinity. In fact, about a half of nsSNPs destabilize/stabilize the complexes by more than 1kcal/mol. In addition, we find that 31.8% of nsSNPs affect protein-protein binding by more than 2kcal/mol and 23.9% by more than 3kcal/mol.

As was mentioned before, in the case of non-OMIM complexes there is no information about which nsSNP is the dominant allele. However, our numerical protocol builds a 3D model of the first allele in the list, minimizes the structure and then introduces a side chain mutation at the nsSNP position and minimizes the mutant structure. Could this bias the calculations? Since $\Delta\Delta\Delta G(nsSNP)$ is a difference between

two binding energies, the change of the order will simply change the sign of the $\Delta\Delta\Delta G(nsSNP)$. If the numerical protocol is not biased, then we should see that the effect of, for example, a P→C mutation is opposite to the effect of a C→P variation. Comparing the means reported in the supplementary results, Table D.1, we can see that this is the case, except for C→H and H→C. (in both cases the means of the distributions of $\Delta\Delta\Delta G_{tot}(nsSNP)$ were found to be negative). However, this is the smallest subset in our study comprised of only 5 cases and many more examples are needed to draw a conclusion.

Another important issue to address is how sensitive the results are in respect to the computational protocol and force field used. Recently we have demonstrated that the calculations of absolute value of the binding energy are very sensitive to both computational protocol and force fields [65]. The same study [65], however, found that the distribution of the binding energy and the general trends are almost insensitive to force field and protocol used. Since the present study is not aimed at computing the absolute binding energy, but rather the change of the binding energy upon single amino acid substitution, the effects of force field and computational algorithm are expected to largely cancel out.

It is expected that a mutation that changes the physico-chemical property of a position at the interface of the corresponding protein-protein complex should affect binding affinity. However, our results indicate that this is not necessarily the case. The outcome of the mutation depends on a variety of factors, whose interplay determines the effects of the substitution. In addition, some positions are located in structural regions

that allow for structural relaxations. From an energetics perspective, an amino acid substitution may not always affect the binding affinity. An example includes a charged residue for which the favorable pair-wise interactions are almost entirely cancelled by an unfavorable desolvation penalty. Another example is weak hydrogen bonds formed at the interface. A third example is a partially exposed hydrophobic residue at the periphery of the interface. Substitution of such residues with another may not affect the binding affinity; in fact, the nsSNP mutation could strengthen the binding.

A highly conserved position within the protein sequence is often related to an important biological function. Multiple sequence alignment analysis showed that most of the positions corresponding to interfacial nsSNPs in our dataset are highly conserved. It was shown that the variance of the total binding energy and its components of the highly conserved positions is larger as compared with the variance of positions with lower conservation. However, significant fraction of nsSNP occurring at conserved positions was calculated not to change the binding energy by much. This indicates that conservation of amino acids in certain interface positions does not occur to preserve binding affinity. Rather, such conservation may reflect the preservation of the binding mode or specificity. An interesting case is an nsSNP mutation that introduces an amino acid found in another species. Since such a mutation was evolutionarily accepted in the other species, the overall effect on protein-protein affinity is expected to be small. In further work, we will explore this observation and will determine the effects of introducing mutations to any other 20 amino acids.

In this paper, we showed that that the change of the binding energy from the target complex to the nsSNP variant is not related to the conservation of the net charge, hydrophobicity or hydrogen bond network. This implies that one cannot simply use the physical-chemical properties of amino acids to evaluate the effects an nsSNP has on protein-protein interactions. Rather detailed structure-based energy calculations must be performed in order to predict these effects, as it was done in the present work.

REFERENCES

1. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A: Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet* 2007, 16(1):1-14.

2. Mooney S: Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform* 2005, 6(1):44-56.

3. Dominy BN: Molecular recognition and binding free energy calculations in drug development. *Curr Pharm Biotechnol* 2008, 9(2):87-95.

4. Huang N, Jacobson MP: Physics-based methods for studying protein-ligand interactions. *Curr Opin Drug Discov Devel* 2007, 10(3):325-331.

5. Jones S, Thornton J: Principles of protein-protein interactions derived from structural studies. *Proceedings of the National Academy of Sciences* 1996, 93:13-20.

6. Vajda S, Vakser I, Steinberg M, Janin J: Modeling of Protein Interactions in Genomes. *Proteins* 2002, 47:444-446.

7. Aloy P, Russell RB: Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* 2006, 7(3):188-197.

8. Gilson MK, Zhou HX: Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* 2007, 36:21-42.

9. Alexov E: Protein-protein interactions. *Curr Pharm Biotechnol* 2008, 9(2):55-56.

10. Villoutreix BO, Bastard K, Sperandio O, Fahraeus R, Poyet JL, Calvo F, Deprez B, Miteva MA: In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics. *Curr Pharm Biotechnol* 2008, 9(2):103-122.

11. Kuntz ID: Structure-Based strategies for drug design and discovery. *Science* 1992, 257:1078.

12. Kick E, Roe D, Skillman A, Liu G, Ewing T, Sun Y, Kuntz I, Ellman J: Structure-based design and combinatorial chemistry yield low nanomolar constants of cathepsin D. *Chem Biol* 1997, 4:297-307.

13. Cavasotto CN, Orry AJ, Abagyan RA: Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* 2003, 51(3):423-433.

14. Gonzalez-Ruiz D, Gohlke H: Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr Med Chem* 2006, 13(22):2607-2625.

15. Teng S, Michonova-Alexova E, Alexov E: Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Curr Pharm Biotechnol* 2008, 9(2):123-133.

16. Koukouritaki SB, Poch MT, Henderson MC, Siddens LK, Krueger SK, VanDyke JE, Williams DE, Pajewski NM, Wang T, Hines RN: Identification and functional analysis of common human flavin-containing monooxygenase 3 genetic variants. *J Pharmacol Exp Ther* 2007, 320(1):266-273.

17. Ode H, Matsuyama S, Hata M, Neya S, Kakizawa J, Sugiura W, Hoshino T: Computational characterization of structural role of the non-active site mutation M36I of human immunodeficiency virus type 1 protease. *J Mol Biol* 2007, 370(3):598-607.

18. De Cristofaro R, Carotti A, Akhavan S, Palla R, Peyvandi F, Altomare C, Mannucci PM: The natural mutation by deletion of Lys9 in the thrombin A-chain affects the pKa value of catalytic residues, the overall enzyme's stability and conformational transitions linked to Na+ binding. *Febs J* 2006, 273(1):159-169.

19. Shirley BA, Stanssens P, Hahn U, Pace CN: Contribution of Hydrogen Bonding to the Conformational Stability of Ribonuclease T1. *Biochemistry* 1992, 31:725-732.

20. Inoue M, Yamada H, Yasukochi T, Kuroki R, Miki T, Horiuchi T, Imoto T: Multiple role of hydrophobicity if tryptophan-108 in chicken lysozyme: structural stability, saccharide binding ability, and abnormal pKa of glutamic acid-35. *Biochemistry* 1992, 31:5545-5553.

21. Stevanin G, Hahn V, Lohmann E, Bouslam N, Gouttard M, Soumphonphakdy C, Welter ML, Ollagnon-Roman E, Lemainque A, Ruberg M, Brice A, Durr A: Mutation in the catalytic domain of protein kinase C gamma and extension of the phenotype associated with spinocerebellar ataxia type 14. *Arch Neurol* 2004, 61(8):1242-1248.

22. Sunyaev S, Ramensky V, Bork P: Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000, 16(5):198-200.

23. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F: SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res* 2005, 33(Database issue):D527-532.

24. Pfeifer D, Pantic M, Skatulla I, Rawluk J, Kreutz C, Martens UM, Fisch P, Timmer J, Veelken H: Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood* 2007, 109(3):1202-1210.

25. Paladini F, Cocco E, Cauli A, Cascino I, Vacca A, Belfiore F, Fiorillo MT, Mathieu A, Sorrentino R: A functional polymorphism of the vasoactive intestinal peptide receptor 1 gene correlates with the presence of HLA-B (*)2705 in Sardinia. *Genes Immun* 2008.

26. Seithel A, Klein K, Zanger UM, Fromm MF, Konig J: Non-synonymous polymorphisms in the human SLCO1B1 gene: an in vitro analysis of SNP c.1929A>C. *Mol Genet Genomics* 2008, 279(2):149-157.

27. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A: The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* 2007, 16(11):2472-2482.

28. Godzik A, Jambon M, Friedberg I: Computational protein function prediction: are we making progress? *Cell Mol Life Sci* 2007, 64(19-20):2505-2511.

29. Vakser IA, Kundrotas P: Predicting 3D structures of protein-protein complexes. *Curr Pharm Biotechnol* 2008, 9(2):57-66.

30. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P: Prediction of deleterious human alleles. *Hum Mol Genet* 2001, 10(6):591-597.

31. Sunyaev SR, Lathe WC, 3rd, Ramensky VE, Bork P: SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* 2000, 16(8):335-337.

32. Dimmic MW, Sunyaev S, Bustamante CD: Inferring SNP function using evolutionary, structural, and computational methods. *Pac Symp Biocomput* 2005:382-384.

33. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J: Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* 2003, 327(5):1021-1030.

34. Cheng TM, Lu YE, Vendruscolo M, Lio P, Blundell TL: Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol* 2008, 4(7):e1000135.

35. Wang Z, Moult J: SNPs, protein structure, and disease. *Hum Mutat* 2001, 17(4):263-270.

36. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A: LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 2005, 21(12):2814-2820.

37. Yue P, Melamud E, Moult J: SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006, 7:166.

38. Ye Y, Li Z, Godzik A: Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput* 2006:439-450.

39. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: CHARMM: A program for macromolecular energy, minimization and dynamic calculations. *J Comp Chem* 1983, 4:187-217.

40. Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH: MMDB: 3D structure data in Entrez. *Nucleic Acids Res* 2000, 28(1):243-245.

41. Altschul S.F. M, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nuclei Acid Res* 1997, 25:3389-3402.

42. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005, 33(Database issue):D514-517.

43. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002, 30(1):52-55.

44. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA: Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* 2000, 15(1):57-61.

45. Shoemaker BA, Panchenko AR, Bryant SH: Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci* 2006, 15(2):352-361.

46. Petrey D, Xiang Z, Tang C, Xie L, Gimpelev M, Mitros T, Soto C, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh I, Alexov E, Honig B: Uning Multiple Structure Alignments, Fast Model Builing, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins* 2003, 53:430-435.

47. Ponder JW: TINKER-software tools for molecular design. In., 3.7 edn. St. Luis: Washington University; 1999.

48. Still WC, Tempczyk A, Hawley RC, Hendrickson T: Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *Journal of the American Chemical Society* 1990, 112:6127-6129.

49. MacKerell Jr. AD, Bashford D, Bellot M, Dunbrack Jr. RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem* 1998, 102:3586-3616.

50. Xiang Z, Honig B: Extending the Accuracy Limits of Prediction for Side-chain Conformations. *J Mol Biol* 2001, 311:421-430.

51. Rocchia W, Alexov E, Honig B: Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J Phys Chem* 2001, 105(85):6507-6514.

52. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B: Rapid Grid-based Construction of the Molecular Surface and the Use of Induced Surface Charges to Calculate Reaction Field Energies: Applications to the Molecular Systems and Geometrical Objects. *J Comp Chem* 2002, 23:128-137.

53. Alexov EG, Gunner MR: Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys J* 1997, 72(5):2075-2093.

54. Georgescu R, Alexov E, Gunner M: Combining Conformational Flexibility and Continuum Electrostatics for Calculating Residue pKa's in Proteins. *Biophysical Journal* 2002, 83:1731-1748.

55. Alexov E: Role of the protein side-chain fluctuations on the strength of pair-wise electrostatic interactions: comparing experimental with computed pK(a)s. *Proteins* 2003, 50(1):94-103.

56. Kundrotas P, Georgieva P, Shosheva A, Christova P, Alexov E: Assesing the quality of the homology-modeled 3D structures from electrostatic standpoint: test on bacterial nucleoside monophosphate kinase families. *J bioinf Comp Biophys* 2007:in press.

57. Zhou N, Wang L: A modified T-test feature selection method and its application on the HapMap genotype data. *Genomics Proteomics Bioinformatics* 2007, 5(3-4):242-249.

58. Neely JG, Hartman JM, Forsen JW, Jr., Wallace MS: Tutorials in clinical research: VII. Understanding comparative statistics (contrast)--part B: application of T-test, Mann-Whitney U, and chi-square. *Laryngoscope* 2003, 113(10):1719-1725.

59. Kowalski CJ, Schneiderman ED, Willis SM: PC program implementing an alternative to the paired t-test which adjusts for regression to the mean. *Int J Biomed Comput* 1994, 37(3):189-194.

60. Brock K, Talley K, Coley K, Kundrotas P, Alexov E: Optimization of electrostatic interactions in protein-protein complexes. *Biophys J* 2007, 93(10):3340-3352.

61. Alexov E: Calculating Proton Uptake/Release and the Binding Free Energy Taking into Account Ionization and Conformation Changes Induced by Protein-Inhibitor Association. Application to Plasmepsin, Cathepsin D and Endothiapepsin-Pepstatin Complexes. *Proteins* 2004, 56:572-584.

62. Alexov E, Miksovska J, Baciou L, Schiffer M, Hanson D, Sebban P, Gunner M: Modeling the Effects of Mutations on the Free Energy of the First Electron Transfer from Qa- to Qb in Photosynthetic Reaction Centers. *Biochemistry* 2000, 39:5940-5952.

63. Alexov E, Gunner M: Calculated Protein and Proton Motions Coupled to Electron Transfer: Electron Transfer from QA- to QB in Bacterial Photosynthetic Reaction Centers. *Biochemistry* 1999, 38:8253-8270.

64. Ofiteru A, Bucurenci N, Alexov E, Bertrand T, Briozzo P, Munier-Lehmann H, Gilles AM: Structural and functional consequences of single amino acid substitutions in the pyrimidine base binding pocket of Escherichia coli CMP kinase. *Febs J* 2007, 274(13):3363-3373.

65. Talley K, Ng K, Shroder M, Kundrotas P, Alexov E: On the electrostatic component of the binding free energy. *PMC Biophysics* 2008.

CHAPTER SIX

STRUCTURAL ASSESSMENT OF THE EFFECTS OF AMINO ACID SUBSTITUTIONS ON PROTEIN STABILITY AND PROTEIN-PROTEIN INTERACTION[5]

ABSTRACT

A structure-based approach is described for predicting the effects of amino acid substitutions on protein function. Structures were predicted using a homology modelling method. Folding and binding energy differences between wild-type and mutant structures were computed to quantitatively assess the effects of amino acid substitutions on protein stability and protein–protein interaction, respectively. We demonstrated that pathogenic mutations at the interaction interface could affect binding energy and destabilise protein complex, whereas mutations at the non-interface might reduce folding energy and destabilise monomer structure. The results suggest that the structure-based analysis can provide useful information for understanding the molecular mechanisms of diseases.

INTRODUCTION

Revealing the effects of amino acid substitutions on protein structure and function is critical for understanding the complex mechanisms of human disease caused by single amino acid mutations. There are 67,000 - 200,000 non-synonymous Single Nucleotide Polymorphisms (nsSNPs) in the human population [1], which give rise to a large number of amino acid substitutions in proteins. The residue changes at key sites within a protein

[5]Teng S, Srivastava AK, Schwartz CE, Alexov E, Wang L: Structural assessment of the effects of amino acid Substitutions on protein stability and protein-protein interaction, Int. J. Computational Biology and Drug Design 2010, 3(4):334-349.

may result in a series of conformation changes, including the breakage of salt bridges, alteration of interaction network, disruption of hydrogen bonds, which in turn may perturb the energy landscape. These changes can affect the kinetics of protein folding or cause protein aggregation and destabilisation [2]. More than half of monogenic diseases are caused by single mutations, and a common mechanism by which amino acid substitutions cause human disease is protein stability change. Yue and Moult investigated the effect of amino acid substitutions on protein stability, and estimated that approximately 25% of nsSNPs in the human population might be deleterious to protein function [3]. Of the known disease-causing missense mutations, the majority (83%) resulted in alternation of protein stability [4].

Amino acid substitutions can also affect protein-protein interactions. Approximately 88% of disease-associated nsSNPs are found to be located in the voids/pockets important for protein-protein interactions [5]. The amino acid substitutions located at the binding interface or active site cleft could block the entrance to the active site, change the recognition, alter the specificity, or affect the binding affinity. For example, the substitution G2019S in leucine-rich repeat kinase 2 (LRRK2) was shown to be associated with familial and sporadic Parkinson's disease [6]. Structure analysis indicates that this mutation is located at the interface of LRRK2's N-terminal and C-terminal domains which is important for positioning of Mg^{2+} within the active site of the kinase [7, 8]. This finding is in agreement with the experimental result that G2019S enhances kinase activity *in vitro* [9]. Recently, Teng et al. [10] examined the effects of nsSNPs at the interaction interfaces of 264 protein complexes using a homology

modeling method and all atoms energy calculations. The results suggest that disease-causing mutations tend to destabilise protein-protein interactions. Therefore, understanding how amino acid substitutions affect protein stability and protein-protein interactions can provide new insights into the molecular mechanisms of human genetic diseases.

Protein structure modeling methods have been widely used for predicting the effects of disease-causing mutations on protein stability and protein-protein interaction. For instance, to predict the effects of the mutations related to the genetic disorder galactosemia, more than one hundred mutant structures of galactose-1-phosphate uridyltransferase were constructed using the homology modeling method, and the results suggested that most mutations might alter protein stability [11]. By mapping disease-causing mutations onto known three-dimensional protein structures, Dimmic and coworkers [12] have shown that about 70% of the deleterious mutations are located in the structurally and/or functionally important sites. However, the effects of mutations were analyzed statically in these studies. The free energy perturbation (FEP) calculation has been used to quantitatively assess the effects of amino acid substitutions on protein stability. Dixit et al. [13] used the AMBER force field and solvent-accessible surface area solvation methods to calculate the protein stability changes in terms of free energy differences caused by cancer-associated mutations in the RET and MET kinases, and showed that the amino acid substitutions could decrease the thermodynamical stability of the mutant structures. The FEP calculation was also used to assess the protein stability changes upon single amino acid substitutions in membrane proteins [14]. Nevertheless,

these studies on FEP calculation did not take into account the effects of amino acid substitutions on protein-protein interactions.

The advent of high-throughput sequencing technology makes it possible to identify a large number of nsSNPs in the human genome. The dbSNP database, one of the primary data resources for genetic studies, contains the information of more than 23 million human SNPs [15]. The records in the dbSNP database are linked to the Online Mendelian Inheritance in Man (OMIM) database, which contains disease gene information, including genetic polymorphisms, map locations, inheritance patterns and clinical descriptions [16]. Computational analyses provide an efficient way for examining the effects of nsSNPs on protein stability and function, and for identifying potential disease-causing mutations. Ng and Henikoff [17] used a position-specific scoring matrix (PSSM) based method called Sorting Intolerant From Tolerant (SIFT) to predict whether an amino acid substitution affects protein function. We have recently developed the MuStab web server for predicting protein stability changes upon amino acid substitutions from sequence features [18]. MuStab uses a support vector machine (SVM) model to discriminate between destabilizing and stabilizing amino acid substitutions in proteins. iPTREE-STAB [19] and I-Mutant 3.0 sequence version [20] are also available for sequence-based prediction of protein stability changes caused by point mutations. Structure-based methods, including PoPMuSiC-2.0 [21], Dmutant [22], Eris [23], I-Mutant 3.0 structure version [20] and FoldX [24], are available for examining the effects of mutations on protein stability and protein-protein interactions. In particular, the FoldX software tool can be used to provide quantitative estimations about the effects of amino

acid substitutions on the stability of proteins or protein complexes using the empirical force field calculation [24]. Among these protein stability predictors, I-Mutant3.0 structure version, Dmutant and FoldX gave the best predictive performances [25].

The experimental approach for determining the effects of amino acid substitutions on protein stability is to obtain the mutant proteins and measure their thermal stability changes by melting experiments. However, the experimental approach is time-consuming and thus may not be applied to a large number of amino acid substitutions. In the present study, a structure-based approach was performed for predicting the effects of amino acid substitutions on protein stability and protein-protein interaction. The differences of folding energy and binding energy between the wild-type and mutant structures were calculated to predict the protein stability and protein-protein interaction changes caused by the mutations. The predictions were evaluated by using other bioinformatic methods. The results suggest that the structure-based approach can provide useful information for characterizing disease-causing mutations in human genetic studies.

METHODS

The schematic diagram of the structure-based approach is shown in Figure 6.1. The methodology was also investigated in two previous studies [26, 27]. For a specific gene with mutations, the related sequence and disease information were extracted from the dbSNP and OMIM databases. If the structure of the target protein was available in the Protein Data Bank (PDB), no structure modeling was needed. Otherwise, target structures were constructed using the homology modeling method [28]. The suitable templates were

identified in the PDB database using the PSI-BLAST program [29], and then used to construct the target structures with the NEST program [30]. Energy minimization was performed to obtain the optimal structure with the TINKER program [31], and the mutant structure was constructed using the SCAP program [32]. The folding energy of the wild-type or mutant structure was calculated using TINKER to estimate the effects of the mutations on protein stability. For amino acid substitutions located at the interface, the binding energy changes were also computed to predict the effects of the mutations on protein-protein interaction. At the end, the predictions were compared with several bioinformatics tools, including FoldX [24], PoPMuSiC-2.0 [21], Dmutant [22], Eris [23], MuStab [18], iPTREE-STAB [19] and I-Mutant 3.0 (both sequence and structure versions) [20].

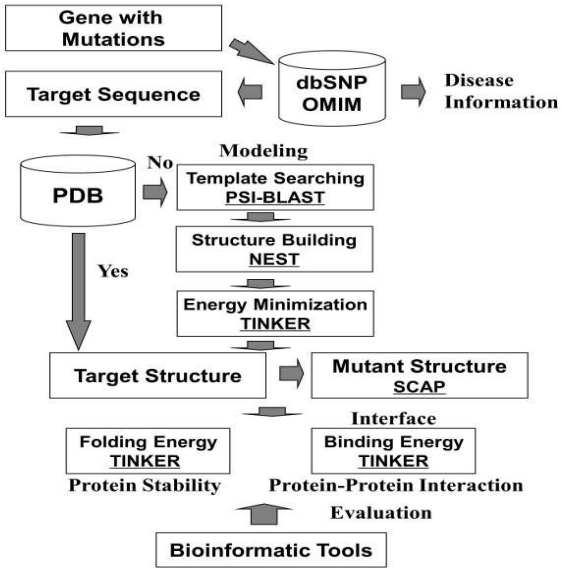

Figure 6.1 Schematic diagram of the approach for assessing the effects of amino acid substitutions on protein stability and protein-protein interaction. Underlined are the software tools used in this study.

Protein structure modeling

Homology modeling was applied to the proteins with no structures available in the PDB database. The structures were modeled as follows:

1) Template searching: The suitable templates were selected from the PDB database for the target protein. Position-Specific Iterated BLAST (PSI-BLAST) [29] was used for the template searching. The structures with significant E-value ($< 10^{-5}$) were selected as the suitable templates.

2) Structure building: The program NEST was used to the build structure models according to the sequence alignment between the target protein and its structural template [30]. NEST is an integrated model-building program, including the program LOOPY9 for loop prediction and SCAP10 for side-chain modeling.

3) Energy minimization: To generate the optimal structure, energy minimization was performed by using the TINKER package [31] with the CHARMM27 force field parameters [33]. The MINIMIZE program in TINKER was used to minimize structures with the algorithm of Limited Memory BFGS Quasi-Newton Optimization [31].

The mutant structures were derived *in silico* from the wild-type structure using the SCAP program [32]. The amino acid substitutions were introduced by side-chain replacements with the rest of the structure kept rigid. The MINIMIZE program in the TINKER package was used to minimize the mutant structures.

Folding energy calculation

The effects of amino acid substitutions on protein stability were assessed by the folding energy changes. The energy calculation was based on the monomer structure of the target protein, and was performed as described in the recent publication [27]. The folding energy is the energy difference between the folded and unfolded states:

$$\Delta G(\,folding\,) = G(\,folded\,) - G(unfolded\,) \qquad (6.1)$$

where $G(folded)$ or $G(unfolded)$ is the total potential energy of the target protein in the folded or unfolded state, respectively.

The protein stability change ($\Delta\Delta G_{stability}$) is the folding energy difference between the wild-type (WT) structure and the structure with the amino acid substitution (AAS). It can be calculated using the following equation:

$$\begin{aligned}\Delta\Delta G_{stability} &= \Delta G(\,folding\!:\!WT) - \Delta G(\,folding\!:\!AAS) \\ &= [G(\,folded\!:\!WT) - G(\,folded\!:\!AAS)] - [G(unfolded\!:\!WT) - G(unfolded\!:\!AAS)]\end{aligned} \qquad (6.2)$$

However, the energy difference between the wild-type and mutant proteins in the unfolded state, *G(unfolded: WT) - G(unfolded: AAS)*, is difficult to calculate. In the present study, we assume that the difference of energy in the unfolded state can be estimated by using the substitution site and its neighboring residues. The total potential energy of the eleven-residue segment (*S*11) with the substitution site in the middle position was used to represent the folding energy of the full-length protein in the unfolded state. Equation (2) can thus be rewritten as:

$$\Delta\Delta G_{stability} = [G(\,folded\!:\!WT\,) - G(\,folded\!:\!AAS)] - [G(\,folded\!:\!WT_{s11}) - G(\,folded\!:\!AAS_{s11})] \qquad (6.3)$$

All of the above total potential energy terms were calculated using the ANALYZE program in the TINKER package. A positive value of $\Delta\Delta G_{stability}$ indicates that the amino acid substitution may make the protein more stable, whereas a negative value of $\Delta\Delta G_{stability}$ suggests that the mutation can destabilise the protein.

Binding energy calculation

For an amino acid substitution located at the interaction interface, the binding energy difference of the protein complex between the wild-type and mutant structures was used to assess the effect of the mutation on protein-protein interaction. As described in the previous study [10], the binding energy was calculated using the rigid body approach, in which the structures of the monomers were kept as they were in the dimer complex. The binding energy, $\Delta\Delta G(binding)$, was the difference between the total potential energy of the dimer complex and the individual monomers:

$$\Delta\Delta G(binding) = \Delta G(folding : complex) - \Delta G(folding : A) - \Delta G(folding : B) \quad (6.4)$$

where $\Delta G(folding: complex)$, $\Delta G(folding: A)$ and $\Delta G(folding: B)$ are the folding free energy values of the dimer complex, monomer A and monomer B, respectively. Since the internal mechanical energy values of the unbound and bound monomers are the same, the energy terms in the unfolded state can be canceled out in equation (4). Thus, the binding free energy can be calculated as below:

$$\Delta\Delta G(binding) = G(folded : complex) - G(folded : A) - G(folded : B) \quad (6.5)$$

where $G(folded: complex)$, $G(folded)$ and $G(folded)$ are the total potential energy values of the dimer complex, monomer A and monomer B in the folded state, respectively.

In this study, the total potential energy was computed using the ANALYZE program in the TINKER package. The effect of an amino acid substitution on protein-protein interaction was assessed by using the binding energy difference between the wild-type (WT) structure and the structure with the amino acid substitution (AAS):

$$\Delta\Delta\Delta G(binding) = \Delta\Delta G(binding : WT) - \Delta\Delta G(binding : AAS) \qquad (6.6)$$

A positive value of the binding energy change ($\Delta\Delta\Delta G_{binding}$) indicates that the amino acid substitution may strengthen the binding affinity and make the protein dimer complex more stable. In contrast, a negative value of $\Delta\Delta\Delta G_{binding}$ suggests that the mutation can weaken the binding affinity and destabilise the dimer complex.

Prediction evaluation

Several bioinformatic tools were used to evaluate the predictive power of the structure-based approach used in this study, and the predictions were considered to be reliable if a consensus was reached by most of the predictors. Sequence-based prediction of the direction of protein stability change could give useful information. Three sequence-based tools were used to predict the directions of protein stability changes caused by amino acid substitutions from primary sequence data, including iPTREE-STAB (http://210.60.98.19/IPTREEr/iptree.htm), MuStab (http://bioinfo.ggc.org/mustab/) and I-Mutant3.0 (sequence version) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi).

Structure-based prediction methods could provide quantitative assessment of the effects of amino acid substitutions on protein stability. Khan and Vihinen [25] compared

the predictive performances of different protein stability predictors, and showed that three structure-based tools, including I-Mutant3.0 (structure version) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi), Dmutant (http://sparks.informatics.iupui.edu/hzhou/mutation.html) and FoldX [24] were the most reliable predictors. These three tools were used in this study. Other two structure-based predictors, including PoPMuSiC-2.0 (http://babylone.ulb.ac.be/popmusic/) and Eris (http://eris.dokhlab.org), were also used to calculate the folding energy for monomer structures, respectively. The difference of the folding energy between the wild-type and mutant structures was used to assess the protein stability change caused by an amino acid substitution, and compared with the $\Delta\Delta G_{stability}$ value calculated using the approach applied in this paper. Furthermore, FoldX was also used to determine the interaction energy of complex protein. The effect of an amino acid substitution on protein-protein interaction was estimated by the interaction energy difference of the protein complex between the wild-type and mutant structures ($\Delta\Delta\Delta G_{FoldX}$), which was compared with $\Delta\Delta\Delta G_{binding}$ computed using the method utilized in this study.

In addition, ClustalX [34] was used to perform the multiple sequence alignment for conservation analysis. Protein sequences from different species were downloaded from the NCBI Entrez database using the GENE search option with the gene name as the query.

RESULTS AND DISCUSSION

To evaluate the usefulness of the structure-based approach utilized in this paper, three case studies were carried out for four pathogenic mutations and one neutral nsSNP in three human genes (Table 6.1). One disease-causing mutation, A111V (dbSNP ID: rs28928889, OMIM ID: 141850.0029), and one neutral nsSNP, T119N (dbSNP ID: rs1058069), in the human *HBA2* gene (haemoglobin subunit alpha) were used to show their different effects on protein stability and protein-protein interaction. Two pathogenic mutations, Q61K (dbSNP ID: rs28933406, OMIM ID: 190020.0002) and A146T (dbSNP ID: rs104894231, OMIM ID: 190020.0008), in the human *HRAS* gene (v-Ha-ras Harvey rat sarcoma viral oncogene homolog) were analyzed to assess the effects of mutations on different structural regions (interface or non-interface). The computational approach was also used to investigate the substitution, A693V, in the human *ZBTB20* gene (zinc finger and BTB domain containing 20). As discussed in the following sections, the results suggest that the pathogenic mutations make the monomer structures less stable ($\Delta\Delta G_{stability} < 0$), and/or weaken the binding affinity to destabilise the dimer structures ($\Delta\Delta\Delta G_{binding} < 0$). In contrast, the neutral nsSNP has only slight effects on protein stability and protein-protein interaction ($\Delta\Delta G_{stability}$ and $\Delta\Delta\Delta G_{binding}$ close to 0).

It was shown that the predictions agree well with the results gave by the most of structure-based methods. However, the sequence-based tools often did not agree with the consensus predictions from the structure-based methods (Table 6.1). The structure-based predictors (I-Mutant3.0 structure version, Dmutant and FoldX) appeared to be more

reliable for predicting protein stability changes caused by mutations [25]. Thus, this study focused on the structure-based analyses.

Table 6.1 The effects of five amino acid substitutions on protein stability. The unit of energy change is kcal/mol.

Amino acid substitution		A111V (HBA2)	T119N (HBA2)	Q61K (HRAS)	A146T (HRAS)	A693V (ZBTB20)
Structure-based Tools	$\Delta\Delta G_{stability}$	-0.75	0.06	-4.42	-1.39	-2.69
	FoldX	-4.19	10.54	-2.74	-0.22	-0.68
	PoPMuSiC-2.0	-0.49	-0.50	-0.24	-0.38	-0.05
	Dmutant	-0.48	0.32	-0.38	-0.24	-0.34
	Eris	4.28	2.29	-2.63	-1.24	-0.72
	I-Mutant 3.0 (structure version)	0.13	-0.30	0.29	-0.79	-0.03
	Consensus	Decreased	Increased	Decreased	Decreased	Decreased
Sequence-based Tools	**I-Mutant 3.0 (sequence version)**	Increased	Increased	Increased	Decreased	Increased
	MuStab	Increased	Decreased	Increased	Decreased	Increased
	iPTREE-STAB	Increased	Decreased	Increased	Decreased	Decreased
	Consensus	Increased	Decreased	Increased	Decreased	Increased

Pathogenic mutation and neutral nsSNP in haemoglobin

Haemoglobin molecules in red blood cells transport oxygen from the lung to the peripheral tissues, and thus are important for maintaining cell viability. Human haemoglobin is made up of symmetric dimers of polypeptide chains, the α/β-globin dimers [35]. Several point mutations in α-globin have been shown to cause α-thalassemia, which can result in Hydrops fetalis [36]. In this study, the two amino acid substitutions of human haemoglobin subunit alpha (HBA2), A111V and T119N, were analyzed to show the different effects of disease-causing and neutral amino acid substitutions on protein

stability and protein-protein interaction. The homodimer structure of HBA2 was built using the crystal structure of human deoxy haemoglobin (PDB: 1O1P) as the template.

The majority of disease-causing mutations cause protein destabilisation, whereas most neutral nsSNPs have limited effect on protein stability [4]. In the present study, the predicted effects of A111V (disease-causing) and T119N (neutral) on protein stability agree well with the previous observations. As shown in Table 6.1, the folding energy change ($\Delta\Delta G_{stability}$) caused by A111V is -0.75 kcal/mol, suggesting that the mutation may destabilise haemoglobin monomer structure. The decreased protein stability is also predicted for the A111V mutation by three structure-based tools including FoldX, PoPMuSiC-2.0 and Dmutant (Table 6.1). In contrast, the neutral nsSNP (T119N) is predicted by our calculations and three structure-based tools (FoldX, Dmutant and Eris) to stabilize the protein monomer. PoPMuSiC-2.0 and I-Mutant3.0 (structure version) give the opposite predictions. The results suggest that T119N may not cause destabilisation of the monomer structure.

Amino acid substitutions at the interaction interface may result in binding affinity changes, and thus affect the structure of the protein complex. As shown in Figure 6.2a, the pathogenic mutation, A111V, is located in the α-helix of the HBA2 binding interface. Although most regions of the wild-type and mutant structures are similar, the structures are not overlapped in the α-helix interface region. This structural change may significantly affect the binding energy, and make the protein complex unstable. The observation has been confirmed by the binding energy calculation using both TINKER and FoldX ($\Delta\Delta\Delta G_{binding}$ = -11.56 kcal/mol and $\Delta\Delta\Delta G_{FoldX}$ = -1.41 kcal/mol) (Table 6.2).

In contrast, the neutral nsSNP (T119N) is located in the flexible loop region (Figure 6.2b). Since T119N is not located in the inner region of the interface, it may not significantly affect protein-protein interaction. The binding energy change caused by T119N is $\Delta\Delta\Delta G_{binding}$ = 0.90 kcal/mol (Table 6.2), which is smaller than the absolute value of binding energy change caused by A111V.

In addition, the multiple sequence alignment shown in Figure 6.2c suggests that the residue, Ala 111, is well conserved, but Thr 119 is not conserved in *Xenopus laevis* and *Xenopus tropicalis*. The result agrees with the previous observation that pathogenic mutations tend to be located at evolutionarily conserved positions [37].

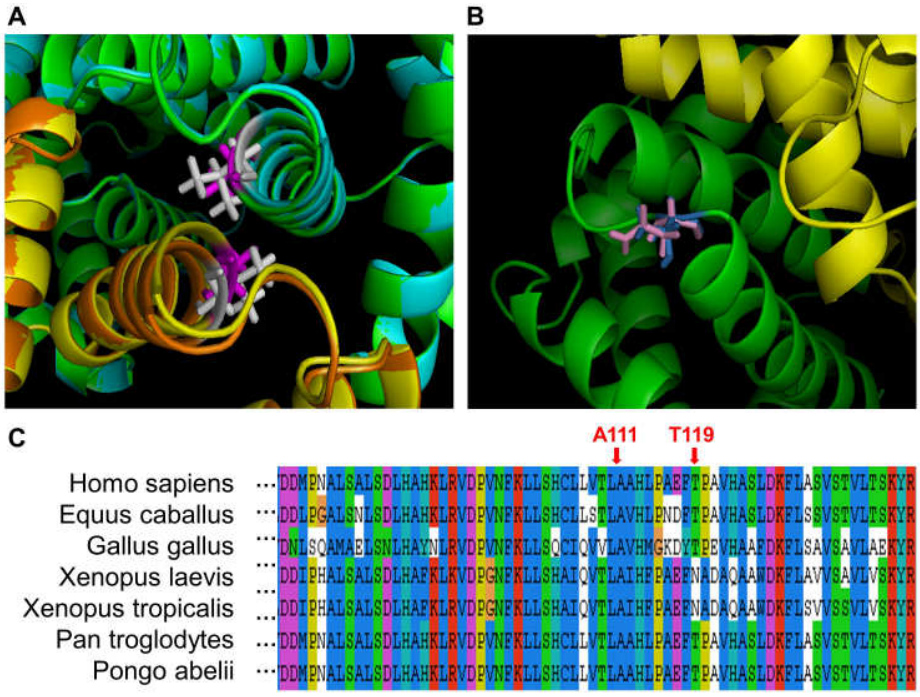

Figure 6.1 Illustration of two amino acid substitutions (A111V and T119N) in human HBA2. a) Structural representation of the A111V mutation. The wild-type chain A is shown in green color, mutant chain A in cyan, wild-type chain B in yellow, and mutant chain B in orange. The amino acid residue Ala 110 (wild-type) is shown in magenta, and Val 110 (mutant) in white. b) Structural representation of T119N. Chains A and B are shown in green and yellow, respectively. The residue Asn 119 (wild-type) is shown in pink, and Thr 119 (mutant) in blue. c) Multiple sequence alignment of HBA2 with the amino acid substitution sites indicated.

Table 6.2 The effects of five amino acid substitutions on protein-protein interaction. The unit of energy change is kcal/mol.

Amino acid substitution	A111V (HBA2)	T119N (HBA2)	Q61K (HRAS)	A146T (HRAS)	A693V (ZBTB20)
$\Delta\Delta G_{binding}$	-11.56	0.90	-7.29	-0.21	-0.31
$\Delta\Delta G_{FoldX}$	-1.41	1.39	-2.40	-0.11	0.00

Pathogenic mutations at the interface or non-interface of HRAS

Follicular carcinoma is the second most common thyroid cancer, which accounts for about 15% of all thyroid malignancies. The v-Ha-ras Harvey rat sarcoma viral oncogene homolog (*HRAS*) encodes a follicular cancer-related protein located at the inner surface of cell membrane. The protein plays an important role in the transduction of signals arising from tyrosine kinase and G protein-coupled receptors. One pathogenic mutation (Q61K) in *HRAS* was found to cause constitutive activation of the downstream signaling pathways [38]. Another disease-causing mutation (A146T) was identified in patients with Costello syndrome, and was shown to affect the GTP/GDP binding of HRAS [39]. In this study, the heterodimer structure of HRAS has been built using the crystal structure of the transforming protein RhoA (PDB: 1OW3) as the template. The amino acid substitution Q61K is located at the interaction interface (Figure 6.3a), and A146T lies in a non-interface region of HRAS (Figure 6.3b). These two mutations in different structural regions have been analyzed to assess their effects on protein stability and protein-protein interaction.

Both amino acid residues in the HRAS protein, Gln 61 and Ala 146, are conserved in other species (Figure 6.3c), suggesting that they may be functionally important sites. As shown in Table 6.1, the folding energy changes ($\Delta\Delta G_{stability}$) caused by Q61K and A146T are -4.42 kcal/mol and -1.39 kcal/mol (Table 6.1), respectively, suggesting that both mutations may destabilise the HRAS monomer structure. Consistent with the above results, the predictions made by structure-based tools show decreased protein stability for both mutations (excluding I-Mutant3.0 structure version for Q61K).

Furthermore, all the sequence-based methods also predict that A146T could make HRAS protein unstable.

The Q61K mutation is located at the interaction interface (Figure 6.3a), and the binding energy change caused by Q61K is $\Delta\Delta\Delta G_{binding}$ = -7.29 kcal/mol, or $\Delta\Delta\Delta G_{FoldX}$ = -2.40 kcal/mol (Table 6.1), suggesting that the mutation may significantly affect protein-protein interaction. The distance between Gln 61 and its interaction partner, Arg 47 from the other chain, is only 1.88 Å, which is within the distance of hydrogen bond formation. When the polar residue Gln is replaced by positively charged residue Lys, the hydrogen bonds may be affected, and thus make strongly unfavorable interactions with Arg 47. In contrast, the A146T mutation located in a non-interface region (Figure 6.3b) does not appear to have a significant effect on protein-protein interaction. As shown in Table 6.1, the binding energy change caused by A146T is $\Delta\Delta\Delta G_{binding}$ = -0.21 kcal/mol, or $\Delta\Delta\Delta G_{FoldX}$ = -0.11 kcal/mol. Nevertheless, Ala 146 and its neighboring residues (Leu 15 and Val 148) may form the hydrophobic pocket, which is involved in the binding of the purine ring of GTP/GDP. The substitution of Ala 146 by the polar residue Thr may alter the hydrophobic environment in the pocket, and thus affect the binding of GTP or GDP.

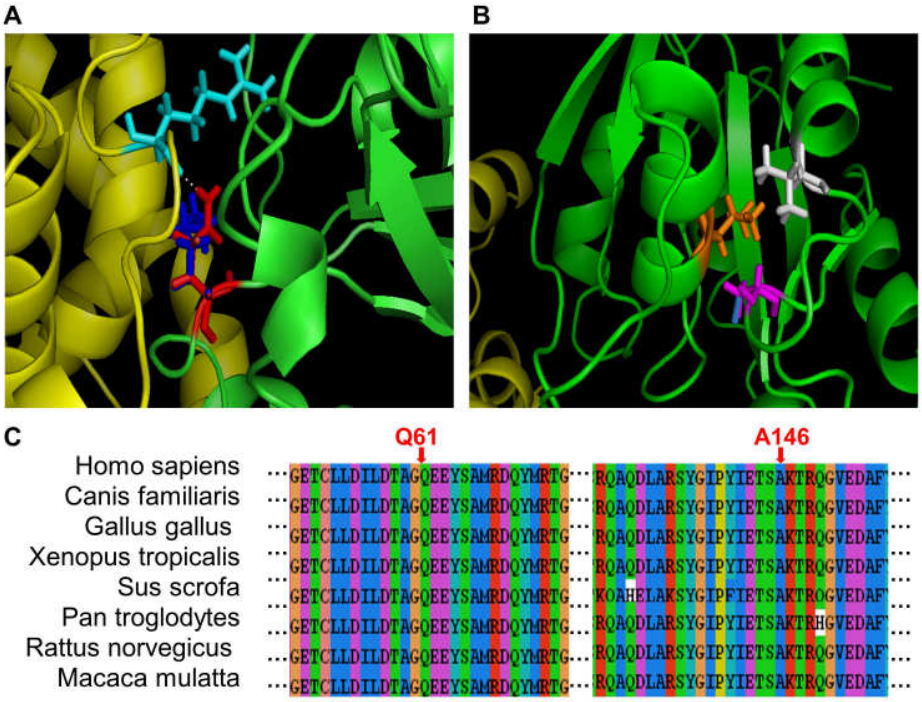

Figure 6.3 Illustration of two disease-causing mutations (Q61K and A146T) in human HRAS. a) Structural representation of the Q61K mutation. Chains A and B are shown in green and yellow, respectively. The residue Gln 61 (wild-type) is shown in red, Lys 61 (mutant) in blue, and Arg 47 of chain B in cyan. The hydrogen bond is represented as a white dash line. b) Structural representation of the A146T mutation. Chains A and B are shown in green and yellow, respectively. Ala 146 (wild-type) is shown in magenta, and Thr 146 (mutant) in blue. Two neighboring residues, Leu 15 in orange and Val 148 in white, are also shown. c) Multiple sequence alignment of HRAS with the amino acid substitution sites indicated.

Application: the A693V substitution in ZBTB20

The structure-based approach was also used to investigate the amino acid substitution, A693V, in the human *ZBTB20* gene (zinc finger and BTB domain containing 20). *ZBTB20* plays important roles in neurogenesis [40], postnatal survival and glucose homeostasis [41]. The A693V substitution is implicated to impair the function of ZBTB20 in the brain. Thus, predicting the effects of A693V on protein

stability and function may help determine the pathogenic potential of the amino acid substitution.

The structure of the C-terminal region (560-739) of ZBTB20, including five zinc finger domains, was constructed using the homology modeling method with the six-finger zinc finger peptide (PDB: 2I13) as the template. As shown in Figure 6.4a, although ZBTB20 may form a homodimer structure, the A693V mutation is not located at the interaction interface. The binding energy change caused by A693V is $\Delta\Delta\Delta G_{binding}$ = -0.31 kcal/mol, or $\Delta\Delta\Delta G_{FoldX}$ = 0 kcal/mol (Table 6.2), suggesting that the amino acid substitution has little effect on dimer formation. The folding energy change was also calculated for A693V using TINKER ($\Delta\Delta G_{stability}$ = -2.69 kcal/mol, Table 6.1). In addition, all of the structure-based methods predicted that A693V will decrease protein stability. Thus, the consensus prediction is that A693V will slightly destabilise the monomer structure of ZBTB20.

Since the ZBTB20 protein was previously shown to bind DNA [40], the structure of ZBTB20 in complex with DNA has been modeled using the six-finger zinc finger peptide (PDB: 2I13) as the template. As shown in Figure 6.4b, the amino acid residue, Ala 693, is located close to the phosphate group of DNA backbone. Therefore, another possibility is that the A693V substitution may be involved in protein-DNA interaction. The multiple sequence alignment shown in Figure 6.4c also suggests that Ala 693 is highly conserved in other species, and thus may be important for the normal function of ZBTB20

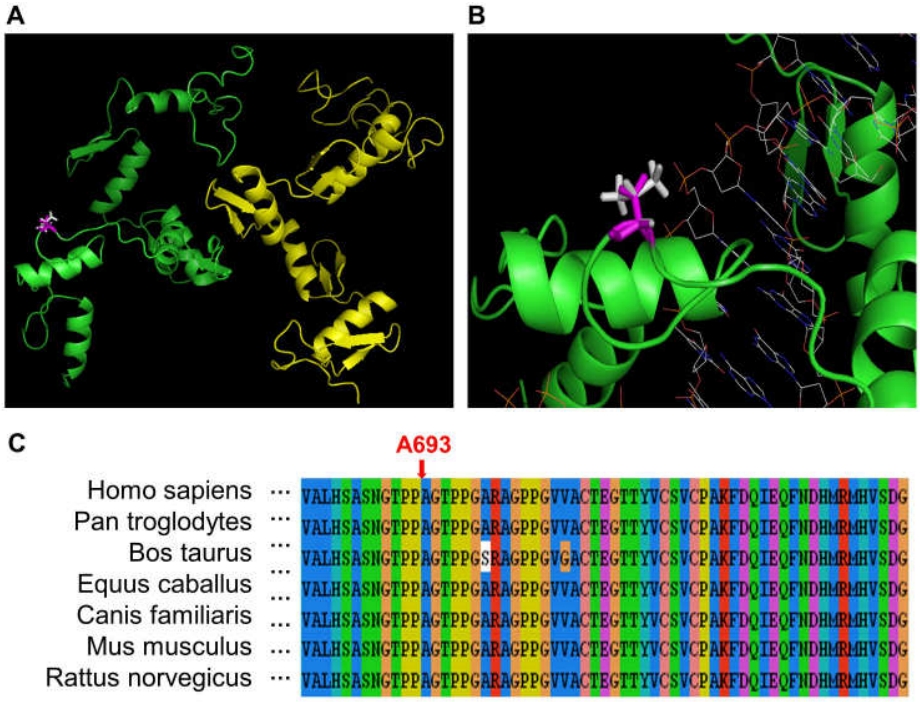

Figure 6.4 Illustration of the A693V mutation in human ZBTB20. a) Structural representation of the A693V mutation. Chains A and B are shown in green and yellow, respectively. Ala 693 (wild-type) is shown in magenta, and Val 693 (mutant) in white. b) Representation of the modeled structure of ZBTB20 in complex with DNA. Shown are chain A in green, Ala 693 in magenta, Val 693 in white, and the DNA molecule as wireframe. c) Multiple sequence alignment of ZBTB20 with the amino acid substitution site indicated.

CONCLUSION

In this paper, a structure-based approach is described for assessing the effects of amino acid substitutions on protein stability and protein-protein interaction. Homology modeling and free energy calculation methods were used to compute the differences of folding energy and binding energy between the wild-type and mutant structures. Three case studies showed that the disease-causing mutations at the interaction interface might reduce the binding energy, and thus weaken the affinity in the protein complex. The

pathogenic mutations in the non-interface region could reduce the folding energy and thus destabilise the monomer structure. Therefore, the structure-based approach can be used to quantitatively assess the effects of amino acid substitutions on protein stability and protein-protein interaction. The approach may be useful for understanding the molecular mechanisms by which gene mutations cause human diseases.

REFERENCES

1. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999, 22(3):231-238.

2. Dill KA, Fiebig KM, Chan HS: Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci U S A* 1993, 90(5):1942-1946.

3. Yue P, Moult J: Identification and analysis of deleterious human SNPs. *J Mol Biol* 2006, 356(5):1263-1274.

4. Wang Z, Moult J: SNPs, protein structure, and disease. *Hum Mutat* 2001, 17(4):263-270.

5. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J: Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* 2003, 327(5):1021-1030.

6. Aasly JO, Toft M, Fernandez-Mata I, Kachergus J, Hulihan M, White LR, Farrer M: Clinical features of LRRK2-associated Parkinson's disease in central Norway. *Ann Neurol* 2005, 57(5):762-765.

7. Albrecht M: LRRK2 mutations and Parkinsonism. *Lancet* 2005, 365(9466):1230.

8. Mata IF, Wedemeyer WJ, Farrer MJ, Taylor JP, Gallo KA: LRRK2 in Parkinson's disease: protein domains and functional insights. *Trends Neurosci* 2006, 29(5):286-293.

9. Kachergus J, Mata IF, Hulihan M, Taylor JP, Lincoln S, Aasly J, Gibson JM, Ross OA, Lynch T, Wiley J, Payami H, Nutt J, Maraganore DM, Czyzewski K, Styczynska M, Wszolek ZK, Farrer MJ, Toft M: Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations. *Am J Hum Genet* 2005, 76(4):672-680.

10. Teng S, Madej T, Panchenko A, Alexov E: Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys J* 2009, 96(6):2178-2188.

11. Facchiano A, Marabotti A: Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. *Protein Eng Des Sel* 2009, 23(2):103-113.

12. Dimmic MW, Sunyaev S, Bustamante CD: Inferring SNP function using evolutionary, structural, and computational methods. *Pac Symp Biocomput* 2005:382-384.

13. Dixit A, Torkamani A, Schork NJ, Verkhivker G: Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: the impact on protein structure, dynamics, and stability. *Biophys J* 2009, 96(3):858-874.

14. Park H, Lee S: Prediction of the mutation-induced change in thermodynamic stabilities of membrane proteins from free energy simulations. *Biophys Chem* 2005, 114(2-3):191-197.

15. Smigielski EM, Sirotkin K, Ward M, Sherry ST: dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000, 28(1):352-355.

16. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2007, 35(Database issue):D5-12.

17. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003, 31(13):3812-3814.

18. Teng S, Srivastava AK, Wang L: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* 2010.

19. Huang LT, Gromiha MM, Ho SY: iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* 2007, 23(10):1292-1293.

20. Capriotti E, Fariselli P, Rossi I, Casadio R: A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* 2008, 9 Suppl 2:S6.

21. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M: Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 2009, 25(19):2537-2543.

22. Zhou H, Zhou Y: Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002, 11(11):2714-2726.

23. Yin S, Ding F, Dokholyan NV: Eris: an automated estimator of protein stability. *Nat Methods* 2007, 4(6):466-467.

24. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: The FoldX web server: an online force field. *Nucleic Acids Res* 2005, 33(Web Server issue):W382-388.

25. Khan S, Vihinen M: Performance of protein stability predictors. *Hum Mutat* 2010, 31(6):675-684.

26. Teng S, Michonova-Alexova E, Alexov E: Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Curr Pharm Biotechnol* 2008, 9(2):123-133.

27. Zhang Z, Teng S, Wang L, Schwartz CE, Alexov E: Computational analysis of missense mutations causing Snyder-Robinson syndrome. *Hum Mutat* 2010, 31(9):1043-1049.

28. Xiang Z: Advances in homology protein structure modeling. *Curr Protein Pept Sci* 2006, 7(3):217-227.

29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.

30. Petrey D, Xiang Z, Tang C, Xie L, Gimpelev M, Mitros T, Soto C, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh I, Alexov E, Honig B: Uning Multiple Structure Alignments, Fast Model Builing, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins* 2003, 53:430-435.

31. Ponder JW: TINKER-software tools for molecular design. In., 3.7 edn. St. Luis: Washington University; 1999.

32. Xiang Z, Honig B: Extending the Accuracy Limits of Prediction for Side-chain Conformations. *J Mol Biol* 2001, 311:421-430.

33. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: CHARMM: A program for macromolecular energy, minimization and dynamic calculations. *J Comp Chem* 1983, 4:187-217.

34. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, 23(21):2947-2948.

35. Kan YW: Molecular biology of hemoglobin: its application to sickle cell anemia and thalassemia. *Schweiz Med Wochenschr Suppl* 1991, 43:51-54.

36. Chui DH, Waye JS: Hydrops fetalis caused by alpha-thalassemia: an emerging health care problem. *Blood* 1998, 91(7):2213-2222.

37. Miller MP, Kumar S: Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 2001, 10(21):2319-2328.

38. Nikiforova MN, Lynch RA, Biddinger PW, Alexander EK, Dorn GW, 2nd, Tallini G, Kroll TG, Nikiforov YE: RAS point mutations and PAX8-PPAR gamma rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma. *J Clin Endocrinol Metab* 2003, 88(5):2318-2326.

39. Zampino G, Pantaleoni F, Carta C, Cobellis G, Vasta I, Neri C, Pogna EA, De Feo E, Delogu A, Sarkozy A, Atzeri F, Selicorni A, Rauen KA, Cytrynbaum CS, Weksberg R, Dallapiccola B, Ballabio A, Gelb BD, Neri G, Tartaglia M: Diversity, parental germline origin, and phenotypic spectrum of de novo HRAS missense changes in Costello syndrome. *Hum Mutat* 2007, 28(3):265-272.

40. Mitchelmore C, Kjaerulff KM, Pedersen HC, Nielsen JV, Rasmussen TE, Fisker MF, Finsen B, Pedersen KM, Jensen NA: Characterization of two novel nuclear BTB/POZ domain zinc finger isoforms. Association with differentiation of hippocampal neurons, cerebellar granule cells, and macroglia. *J Biol Chem* 2002, 277(9):7598-7609.

41. Sutherland AP, Zhang H, Zhang Y, Michaud M, Xie Z, Patti ME, Grusby MJ, Zhang WJ: Zinc finger protein Zbtb20 is essential for postnatal survival and glucose homeostasis. *Mol Cell Biol* 2009, 29(10):2804-2815.

CHAPTER SEVEN

CONCLUSIONS

In the present study, several predictive methods, including machine learning and structure modeling approaches, have been developed for analyzing genes and proteins to discover biological knowledge hidden in the heterogeneous datasets. Machine learning can be used to automatically recognize hidden patterns and make accurate predictions based on models derived from complex data. In this study, machine learning approaches were developed for identification of human tissue-specific genes using microarray gene expression data and sequence-based predictions of protein sumoylation sites and protein stability changes upon amino acid substitutions. The results suggest that the use of relevant biological features for classifier construction can improve the classifier performance, and the approaches and tools developed in the study can provide valuable information for genetic research community. The structure-based approaches were developed to quantitatively assess the effects of amino acid substitutions on protein stability and protein-protein interaction. It has been shown that pathogenic mutations can reduce the folding energy to destabilize the monomer structures and weaken the binding affinity to make the complex structures less stable. The machine learning approaches together with the structure-based methods were used to analyze candidate genes and proteins associated with human genetic disorders such as intellectual disability.

Genes and proteins play essential roles in almost every biological process within the cell. The predictive bioinformatic approaches developed in this study may help understand the molecular mechanisms of tissue-specific gene expression, protein

sumoylation, protein stability and protein-protein interaction. Tissue-specific genes and protein sumoylation targets are implicated in many complex diseases. The amino acid substitutions at protein key sites play important roles in many monogenic diseases. However, molecular mechanisms underlying the genetic disorders are still poorly understood. The methods developed in the present study have been used to analyze disease candidate genes and proteins for human genetic studies and the findings may help elucidate the molecular mechanisms of some genetic disorders such as intellectual disability. The analytical results may also help biomedical scientists to design their experiments and to interpret the experimental data. Furthermore, the web servers make our computational methods available to the broader scientific community.

The difficulty of predicting protein stability changes upon amino acid substitutions with machine learning approaches lies in the rarity of known positive and negative examples. Thus, semi-supervised learning methods, such as Self-training, Co-training, Semi-supervised Support Vector Machines and Graph-based methods, can be used to improve the classifier performance with both labeled and unlabeled data in future works. Moreover, feature selection approaches, such as Random Forests and Support Vector Machine-based Recursive Feature Elimination, can be used to select relevant features for classifier construction to enhance the classifier performance. The predictive methods developed in this study can be used to analyze intellectual disability candidate genes for future studies.

APPENDICES

Appendix A

Publications resulting from the present research

CHARPTER2

Teng S, Wang L: A machine learning approach for predicting human tissue-specific genes using microarray expression data, in preparation.

CHARPTER3

Teng S, Luo H, Wang L: Predicting protein sumoylation sites from sequence features, submitted.

Teng S, Luo H, Wang L: Random Forest-Based Prediction of Protein Sumoylation Sites from Sequence Features. In: *Proceedings of the 2010 ACM-BCB*. Association for Computing Machinery, 2010:120-126.

CHARPTER4

Teng S, Srivastava AK, Wang L: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* 2010, 11(Suppl 2):S5.

Teng S, Srivastava AK, Wang L: Biological Features for Sequence-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. In: *Proceedings of IJCBS'09* IEEE Computer Society, 2009:201-206.

CHARPTER5

Teng S, Kundrotas P, Madej T, Panchenko A, Alexov E: Modeling effects of human SNPs on protein-protein interactions. *Biophysics. J.* 2009, 96(6):2178-2188.

CHARPTER6

Teng S, Srivastava AK, Schwartz CE, Alexov E, Wang L: Structural assessment of the effects of amino acid Substitutions on protein stability and protein-protein interaction, *Int. J. Computational Biology and Drug Design* 2010, 3(4):334-349

OTHER RELATED PUBLICATIONS

Teng S, Michonova-Alexova E, Alexov E: Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Curr Pharm Biotechnol* 2008, 9(2):123-133.

Zhang Z[§], Teng S[§], Wang L, Schwartz CE, Alexov E: Computational analysis of missense mutations causing Snyder-Robinson syndrome. *Human Mutation* 2010, Sep 31(9):1043-9. ([§]These authors contributed equally to this manuscript)

Appendix B

Additional files for predicting human tissue-specific genes

Additional file B1. List of brain-specific gene targets. (Supplemental 1)

Additional file B2. List of liver-specific gene targets. (Supplemental 1)

Appendix C

Additional files for Predicting Protein Sumoylation Sites

Table C.1 The list of 457 experimentally verified sumoylation sites in 263 proteins

Protein Accession (UniProt Entry)	Sumoylation Site Position	Core Motif	Match ΨKXE (Yes/No)	Reference (PMID)
268 sumoylation sites used by SUMOpre (reported before 08/10/2006)				
Q9H3D4	588	LKIP	N	15539951
Q9H3D4	676	IKEE	Y	15539951
O56136	84	EKGE	N	15527853
O56136	447	FKFE	N	15527853
P19544	73	IKQE	Y	15520190
P19544	177	FKHE	N	15520190
O75030	289	IKRE	Y	15507434
O75030	423	IKQE	Y	15507434
P19532	330	IKRE	Y	15507434
P19484	347	VKQE	Y	15507434
P37231	107*	IKVE	Y	15229330
P37231-2	365	PKFE	Y	16127449
Q02447	120	IKDE	Y	12419227
Q02447	551	IKEE	Y	12419227
Q16665	391	LKKE	Y	15465032
Q16665	477	LKLE	Y	15465032
P46060	524	LKSE	Y	15355965
Q14191	496	LKME	Y	15355988
P43694	365	IKTE	Y	15337742
Q08211	76	IKSE	Y	15312759
Q08211	120	LKAE	Y	15312759
P36508	411	VKEE	Y	15280358
Q03188	534	VKSE	Y	15272016
Q03188	721	PKNR	N	15272016
Q03188	746	LKPL	N	15272016
P63279	153	AKKF	N	15272016
P19419	230	LKSE	Y	15210726
P19419	249	VKVE	Y	15210726

P19419	254	PKEE	Y	15210726
Q9Y4L2	244*	VKTE	Y	15208321
Q9Y4L2	263	IKDE	Y	15208321
Q13285	119	FKLE	N	15192080
Q13285	194	IKSE	Y	15192080
P15976	137	LKTE	Y	15173587
P15873	127	LKIE	Y	12226657
P15873	164	TKET	N	12226657
Q05193	376	VKME	Y	15123615
Q60591	684	IKTE	Y	15117942
Q60591	897	VKQE	Y	15117942
P41212	99	TKED	N	12626745
O13066	517*	LKSE	Y	15094046
O60812	237	VKME	Y	15082759
P42858	6	EKLM	N	15064418
P42858	9	MKAF	N	15064418
P42858	15	LKSF	N	15064418
P10275	386	IKLE	Y	12177000
P10275	520	VKSE	Y	12177000
Q13485	113	VKYC	N	12621041
Q13485	159	VKDE	Y	12621041
Q03933	82	VKQE	Y	11278381
P78347	221	VKTE	Y	15016812
P78347	240	VKEE	Y	15016812
P78347	456	VKEE	Y	15016812
P78347	991	IKQE	Y	15016812
Q9BYV9	202	EKEE	N	15060166
Q9BYV9	276	IKSE	Y	15060166
Q9BYV9	421	CKQE	N	15060166
Q9BYV9	580	IKCE	Y	15060166
O00429	38	GKSS	N	14972687
Q9Y6K9	277	AKQE	Y	14651848
Q9Y6K9	309	YKAD	N	14651848
Q9UPW6	233	IKVE	Y	14701874
Q9UPW6	350	VKPE	Y	14701874
P06536	297*	VKTE	Y	14663148
P06536	313*	IKQE	Y	14663148
P25963	21	LKKE	Y	14613580

P42224	703	IKTE	Y	12855578
P43354	91	IKVE	Y	14559918
P43354	577	LKLE	Y	14559918
Q05516	242	VKTE	Y	14527952
Q9NS56	560	KKEE	N	14516784
P08235	89	IKTE	Y	14500761
P08235	399	IKPE	Y	14500761
P08235	428	IKQE	Y	14500761
P08235	494	IKQE	Y	14500761
P08235	953	LKVE	Y	14500761
Q05397	152	VKSD	N	14500712
P06401	388	IKEE	Y	12529333
Q09472	1020	LKTE	Y	12718889
Q09472	1024	IKEE	Y	12718889
P17676	174	LKAE	Y	12810706
P11831	147	IKME	Y	12788062
Q13363	428	VKPE	Y	12679040
P32457	4	LKEE	Y	12149243
P32457	11	IKQD	N	12149243
P32457	30	IKQE	Y	12149243
P32457	63	VKVE	Y	12149243
P32457	287	AKSD	N	12761287
P32457	443	AKLE	Y	12761287
P32457	465	QKSE	N	12761287
Q12216	438	VKNE	Y	12761287
Q12216	446	VKQE	Y	12761287
Q99497	130	AKDK	N	12761214
P23769	222	MKME	Y	12750312
P23769	389	MKKE	Y	12750312
Q924A0	297	VKQE	Y	12727872
Q00613	298	VKEE	Y	11514557
P10242	503	IKQE	Y	12631292
P10242	527	IKQE	Y	12631292
P36956	123	IKEE	Y	12615929
P36956	418	VKTE	Y	12615929
Q12772	464	VKDE	Y	12615929
P16220	285	RKRE	N	12552083
P16220	304	KKKE	N	12552083

Q15788	732	IKLE	Y	12529333
Q15788	774	VKVE	Y	12529333
P49715	161	IKQE	Y	12511558
P55854	11*	VKTE	Y	12506199
P56817	275	LKMD	N	12506199
P11387	103	IKKE	Y	11709553
P11387	117	IKDE	Y	11709553
P11387	153	IKTE	Y	11709553
P11387	328	IKEE	Y	11709553
P11387	436	IKGE	Y	11709553
Q13547	444	VKTE	Y	11960997
Q13547	476	VKEE	Y	11960997
P27540	245	VKKE	Y	12354770
O15169	857	GKVE	N	12223491
O15169	860	EKVD	N	12223491
Q9NSC2	1086	IKTE	Y	12200128
P49716	120	LKRE	Y	12161447
Q9UER7	630	CKKS	N	12150977
Q9UER7	631	KKSR	N	12150977
P04150	277	VKTE	Y	12144530
P04150	293	IKQE	Y	12144530
P04150	703	VKRE	Y	12144530
P06786	1220	IKLE	Y	12086615
P06786	1246	IKKE	Y	12086615
P06786	1277	IKKE	Y	12086615
P05549	10	IKYE	Y	12072434
Q92481	10*	VKYE	Y	12072434
Q15596	239	IKEE	Y	12060666
Q15596	731	IKQE	Y	12060666
Q15596	788	EKEE	N	12060666
P56524	559	VKQE	Y	12032081
Q13569	330*	VKEE	Y	11889051
P04637	386	FKTE	N	11867732
P05627	229	LKEE	Y	16055710
P05627	257	IKAE	Y	16055710
P46061	526*	LKSE	Y	11853669
P23497	297	IKKE	Y	11792325
P15330-2	382	IKTE	Y	11756545

P27782	25	FKDE	N	11731474
P27782	267	VKQE	Y	11731474
P29590	65	AKCP	N	9756909
P29590	160	LKHE	Y	9756909
P29590	487	RKVI	N	9756909
P29590	490	IKME	Y	9756909
O15350	627	IKEE	Y	10961991
Q9H2X6	32	LKIE	Y	15766567
Q9H2X6	1191	AKVN	N	12149243
P32458	412	IKQE	Y	12149243
Q07657	426	IKQE	Y	12149243
Q07657	437	IKTE	Y	12149243
P19893	175	IKQE	Y	10684265
P19893	180	IKPE	Y	10684265
P03116	514	IKAP	N	11005821
P13202	450	VKSE	Y	11602710
Q9M0K4	258	KKQE	N	11581165
P03243	104	VKRE	Y	11553772
Q64127	724	IKQE	Y	11313457
Q64127	742	VKQE	Y	11313457
O00541	517	LKLE	Y	11071894
Q6XA64	802	IKSE	Y	15105549
P03206	12	VKFT	N	11160742
P03209	19	IKKQ	N	15229220
P03209	213	SKTG	N	15229220
P03209	517	VKAL	N	15229220
P61086	13	FKEV	N	15723079
P45448	213	LKLE	Y	15713642
P45448	289	IKSE	Y	15713642
P33244-2	418	IKQE	Y	15713642
P57682	10	VKQE	Y	15684403
P57682	198	IKIE	Y	15684403
Q15744	121	VKEE	Y	15661739
Q16514	19	IKPE	Y	15637059
Q15542	14	VKLE	Y	15637059
P04591	474	QKQE	N	15613319
P00445	18	VKFE	Y	15596868
P00445	69	KKTH	N	15596868

P05750	211	PKEE	Y	15596868
P16649	270	PKEE	Y	15596868
Q07979	322	EKNE	N	15596868
Q07979	328	VKQE	Y	15596868
P04456	60	YKVI	N	15542864
P11978	1128	VKNV	N	15542864
Q04322	498	LKMG	N	15542864
P21538	807	MKTE	Y	15542864
Q14814	439	IKSE	Y	15743823
P41970	162	IKRE	Y	15580297
Q9U1H5	438	IKSE	Y	15788563
Q13422	58	VKVE	Y	15767674
Q13422	241	IKEE	Y	15767674
P06400	720	LKFK	N	15806172
P21063	95	IKIE	Y	15800065
O00180	274	LKKF	N	15820677
P54253	16	KKRE	N	15824120
P54253	194	HKAE	N	15824120
P54253	610	LKID	N	15824120
P54253	697	VKKG	N	15824120
P54253	746	LKFP	N	15824120
P54132	317	SKCL	N	15829507
P54132	331	RKED	N	15829507
P54132	344	SKPE	N	15829507
P54132	347	EKMS	N	15829507
Q8N2W9	35	LKHE	Y	15831457
P42575	77	AKVG	N	15882978
P41161	89	IKRE	Y	15857832
P41161	263	FKQE	N	15857832
P41161	293	IKQE	Y	15857832
P41161	350	VKQE	Y	15857832
O60315	391	IKTE	Y	16061479
O60315	866	IKKE	Y	16061479
P01100	265	LKTE	Y	16055710
O92597	158	VKAE	Y	16014952
P14921	15*	IKTE	Y	16319071
Q90YL1	61	LKKE	Y	16256735
Q90YL1	365	IKTE	Y	16256735

Q8AXX8	52	VKKE	Y	16256735
Q8AXX8	341	VKTE	Y	16256735
P18412	54	VKNE	Y	16306045
Q92793	998	MKTE	Y	16287980
Q92793	1033	VKEE	Y	16287980
Q92793	1056	VKVE	Y	16287980
Q16621	368	TKME	N	16287851
P40381	103	LKWE	Y	16168376
O60016	109	VKKE	Y	16168376
O60016	160	VKEE	Y	16168376
O42934	198	LKWE	Y	16168376
Q8N4C6	1641	LKEE	Y	16154161
Q8N4C6	1680	LKDE	Y	16154161
Q01543	67	VKRE	Y	16148010
Q99683	535	AKQE	Y	16142216
Q99683	1083	LKWE	Y	16142216
Q99683	1114	LKLE	Y	16142216
P55265	418	IKLE	Y	16120648
O00327	259	VKVE	Y	16109848
Q969V6	499	VKEE	Y	16098147
Q969V6	576	VKQE	Y	16098147
Q969V6	624	VKQE	Y	16098147
Q5U0M2	15	IKTE	Y	16862185
Q5U0M2	227	IKQE	Y	16862185
Q86YP4	30	IKME	Y	16738318
Q86YP4	487	AKAE	Y	16738318
Q5VUR2	33	LKME	Y	16738318
O95600	67	IKIE	Y	16617055
Q13426	210	IKQE	Y	16478998
Q06413	391	IKSE	Y	16478538
P10636-8	339	VKSE	Y	16464864
Q9NRA1	314	PKTG	N	16443219
P49792	2571	SKVE	N	16194093
P49792	2592	SKVK	N	16194093
P49792	2650	TKLK	N	16194093
P49792	2723	EKAK	N	16194093
P49792	2725	AKAD	N	16194093
P63165	16	DKKE	N	16194093

P63165	37	FKVK	N	16194093
P63165	39	VKMT	N	16194093
P63165	46	KKLK	N	16194093
P61956	5*	EKPK	N	16194093
P61956	11*	VKTE	Y	16194093
P61956	42*	SKLM	N	16194093
P06876	499*	IKRE	Y	16162816
P06876	523*	IKQE	Y	16162816
P12004	164*	AKDG	N	15931174
O15151	254	IKVE	Y	15907800
O15151	379	IKKE	Y	15907800
Q07666	96	VKME	Y	16568089
139 sumoylation sites (reported between 08/10/2006 to 01/01/2010)				
P46060	8	AKLA	N	15355965
P63279	14	RKAW	N	20424159
P63279	49	KKGT	N	20424159
P41212	11	IKQE	Y	18212042
O00429	532*	DKSS	N	19638400
O00429	535*	SKVP	N	19638400
O00429	558*	GKLI	N	19638400
O00429	568*	TKNV	N	19638400
O00429	594*	LKTS	N	19638400
O00429	597*	SKAE	N	19638400
O00429	606*	EKSK	N	19638400
O00429	608*	SKPI	N	19638400
Q05516	387	SKLG	N	17498654
Q05516	396	MKSE	Y	17498654
P55854	41*	SKLM	N	20029837
Q16621	215	AKPT	N	19966288
Q16621	234	MKIP	N	19966288
Q16621	241	DKIV	N	19966288
P63165	7	AKPS	N	20388717
P63165	25	LKVI	N	20388717
O95365	61	KKLF	N	17595526
P17544	118	IKEE	Y	17264123
Q8VIM5	573	IKQE	Y	17101795
Q13263	554	VKEE	Y	17079232
Q13263	575	TKPV	N	17298944

Q13263	676	LKEE	Y	17298944
Q13263	750	EKLS	N	17298944
Q13263	779	DKAD	N	17298944
Q13263	804	TKFS	N	17298944
P04618	115	TKE	N	17067581
Q5U3M4	341	VKEE	Y	17060459
Q99607	657	IKME	Y	16904644
P48432	247	VKSE	Y	17097055
P00441	76	PKDE	Y	16828461
Q13642	144	PKGE	Y	17509614
Q13642	300	VKAP	N	17509614
P20263	118	VKLE	Y	17496161
Q9LSE2	393	VKEE	Y	17416732
Q9UHF7	1201	VKTE	Y	17391059
P25490	288	IKED	N	17353273
P12757	50	VKKE	Y	17202138
P12757	383	IKQE	Y	17450299
Q14526	333	MKHE	Y	17283066
P46099	56	LKSE	Y	17938210
Q04110	5	IKTE	Y	17888002
Q9Y458	63	PKTE	Y	17846996
Q92993	430	LKSE	Y	17704809
Q92993	451	IKKE	Y	17704809
P11474	14	IKAE	Y	17676930
P11474	403	VKLE	Y	17676930
P49841	292	FKFP	N	18949077
P06748	263	PKVE	Y	17951246
P05455	41	IKLD	N	17646655
P54841	32	VKKE	Y	17548468
P54841	297	VKCE	Y	17548468
Q04887	396*	IKTE	Y	17440973
P08047	16	VKIE	Y	18572193
Q9WVS8	6	LKEE	Y	18467627
Q9WVS8	22	VKAE	Y	18467627
P43268	96	IKKE	Y	18447755
P43268	226	FKQE	N	18447755
P43268	260	IKQE	Y	18447755
P43268	322	IKQE	Y	18447755

P43268	441	LKAE	Y	18447755
P06778	126	KKSA	N	18396468
Q64729	391	MKHF	N	18469808
P17275	240	FKEE	N	18424718
Q01826	744	VKLE	Y	18408014
P48552	756	IKSE	Y	18211901
P48552	1154	IKKE	Y	18211901
Q9CAE3	287	IKVE	Y	18069938
Q9CAE3	693	PKAD	N	18069938
Q9CAE3	770	IKAE	Y	18069938
Q9UPG8	250	IKTE	Y	17551969
Q9UPG8	269	VKEE	Y	17551969
Q9UPG8	356	PKVE	Y	17551969
P04050	1487	VKDE	Y	19384408
Q63120	949	IKKE	Y	19074644
Q8CF90	32	VKKE	Y	19029092
Q9UKL0	294	VKKE	Y	18854179
P33242	119*	FKLE	N	18838537
P33242	194	IKSE	Y	18726511
P37238	107*	IKVE	Y	18832723
Q13887	162	IKTE	Y	18782761
Q13887	209	IKQE	Y	18782761
P05067	662	IKTE	Y	18675254
P05067	670	VKMD	N	18675254
P70671	152	LKDE	Y	18635538
P70434	406	VKLE	Y	18635538
Q15022	72	VKKP	N	18628979
Q15022	73	KKPK	N	18628979
Q15022	75	PKME	Y	18628979
P03120	292	LKGD	N	18619639
Q00653	90	AKIE	Y	18617892
Q00653	298	MKIE	Y	18617892
Q00653	689	LKAG	N	18617892
Q00653	863	VKED	N	18617892
Q71A33	33*	VKKE	Y	20127678
Q99814	394	LKEE	Y	20026589
Q19289-2	460	IKLE	Y	19922876
P35187	621	IKRE	Y	19906698

Q499N1	163	KKKE	N	19850744
Q499N1	168	PKPE	Y	19850744
Q499N1	396	LKME	Y	19850744
Q91ZP3	599	IKEE	Y	19753306
Q91ZP3	629	IKHE	Y	19753306
Q92786	556	IKSE	Y	19706680
P07830	69	LKYP	N	19635839
P07830	285	MKCD	N	19635839
Q9UBK2	184	VKTE	Y	19625249
Q9UKL3	1813	LKSE	Y	19615980
Q07869	185	LKAE	Y	19955185
P17655	390	IKLE	Y	19422794
Q9SJN0	391	LKEE	Y	19276109
A9YTQ3	538	IKME	Y	19251700
A9YTQ3	577	LKTE	Y	19251700
A9YTQ3	660	VKRE	Y	19251700
Q9QWM1	173	MKLE	Y	19125815
Q9QWM1	289*	IKSE	Y	19125815
P02545	201	MKEE	Y	18606848
Q9UQ80	93	LKSD	N	19946338
Q9UQ80	298	AKHE	Y	19946338
O13351	14	DKSA	N	19707600
O13351	30	VKPS	N	19707600
Q9UHP3	99	DKDD	N	19440361
Q99856	398	IKKE	Y	19436740
O88275	63	IKPF	N	19339015
O88275	107	IKVE	Y	19339015
Q12692	126	LKVE	Y	19217407
Q12692	133	SKK	N	19217407
Q9QXZ7	178	AKLE	Y	19186166
Q9QXZ7	315	FKPE	N	19186166
Q9QXZ7	322	LKDP	N	19186166
P32333	101	VKLE	Y	19139279
P32333	109	IKLE	Y	19139279
P35398-2	240	IKPE	Y	19041634
Q61164	74	MKTE	Y	19029252
Q61164	698	VKKE	Y	19029252
Q06710	308	IKQE	Y	18974227

50 sumoylation sites (reported between 01/01/2010 to 06/01/2010)				
Q8R4I1	257	LKST	N	19843541
A2RU29	596	MKSE	Y	20501696
A2RU29	649	VKKE	Y	20501696
A2RU29	650	KKEE	N	20501696
A2RU29	739	VKKE	Y	20501696
A2RU29	793	VKAE	Y	20501696
O35426-2	276	VKIE	Y	20408817
O35426-2	297	VKKE	Y	20408817
Q16666	561	LKTE	Y	20388717
Q9UPN6	18	YKPP	N	20388717
Q9UG01	4	LKHL	N	20388717
Q99518	492	QKQR	N	20388717
Q96QD9	140	RKAN	N	20388717
Q9UBD0	215	VKSA	N	20388717
Q9UBG0	1142	QKPL	N	20388717
Q8IYA6	198	RKPD	N	20388717
O75475	75	RKGF	N	20382164
O75475	250	DKKE	N	20382164
O75475	254	GKKE	N	20382164
O75475	364	LKID	N	20382164
O43290	94	VKRE	Y	20346425
O43290	141	IKKE	Y	20346425
P40337	171	VKPE	Y	20300531
P20193	839	IKVF	N	20299342
P31669	469	IKQE	Y	20228053
P06730	36	IKHP	N	20228053
P06730	49	FKND	N	20228053
P06730	162	DKIA	N	20228053
P06730	206	TKSG	N	20228053
P06730	212	TKNR	N	20228053
O14776	503	IKEE	Y	20215116
O14776	608	IKEE	Y	20215116
P58012	114	IKVP	N	20209145
P58012	150	MKRP	N	20209145
Q92585	217	LKQE	Y	20203086
Q92585	299	IKTE	Y	20203086
P43351	411	MKKR	N	20190268

P43351	412	KKRK	N	20190268
P43351	414	RKYD	N	20190268
P08069	1055	MKEF	N	20145208
P08069	1130	NKFV	N	20145208
P08069	1150	VKIG	N	20145208
P54843	33	VKKE	Y	20127678
Q15843	27	IKER	N	20029837
Q15843	33	EKEG	N	20029837
Q15843	54	EKTA	N	20029837
Q12306	54	AKRQ	N	20029837
P17844	53	PKFE	Y	19995069
Q92841	50	PKFE	Y	19995069
Q08499-6	323	VKTE	Y	20196770

* Redundant protein sumoylation site

Table C.2 List of 40 biological features used in chapter three

Class	**Feature**	**Abbreviation**
Biochemical	Hydrophobicity	H
	pKa value	K
	Molecular weight	M
	Size	S
	Residue volume	V
	Polarity	P
	Amino acid composition	Co
	Buriability	Br
	Side chain hydropathy, corrected for salvation	Ss
	Scaled side chain hydrophobicity values	Hs
Structural	Conformational parameter for alpha helix	A
	Conformational parameter for beta sheet	B
	Conformational parameter for beta-turn	T
	Conformational parameter for coil	C
	Buried average area	Aa
	Bulkiness	Bu
	Average accessible surface area	As

Thermodynamic	Entropy of formation	E
	Transfer free energy	Et
	Partition energy	Ep
	Short and medium range non-bonded energy per residue	Er
	Average non-bonded energy per residue	En
	Free energy in alpha-helical conformation	Ea
	Free energy in beta-strand conformation	Eb
	Solvation free energy	Es
	Hydration free energy	Eh
Empirical	Stability scale from the knowledge-based atom-atom potential	S1
	Relative stability scale extracted from mutation experiments	S2
	Side-chain contribution to protein stability (kJ/mol)	S3
	Molar fraction of 3220 accessible residues	Fa
	Atomic weight ratio of hetero elements in end group to C in side chain	Rh
	Mobility on chromatography paper	Mc
Other	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins	Is
	Side chain volume	Vs
	Flexibility	F
	Refractivity	R
	Number of codons	No
	Recognition factors	Rf
	Relative mutability	Rm
	Transmembrane tendency	Tt

Appendix D

Additional data regarding the effect of nsSNPs on binding energy with respect to amino acid characteristics

Four different classes of amino acids were considered based on the amino acids' physico-chemical properties: polar (S, T, H, N, Q, Y), charged (E, D, K, R), hydrophobic (W, I, L, M, F) and small (P, A, G, C, V). Below we illustrate the effects of nsSNP mutations on the $\Delta\Delta\Delta G_{tot}(snSNP)$, $\Delta\Delta\Delta G_{vdw}(nsSNP)$ and $\Delta\Delta\Delta G_{el}(nsSNP)$ (see main body of the manuscript for definition of these quantities) separately for each class and between classes (Table E.1). The corresponding distributions are shown in Figure E.1. It can be seen that mutations that replace either charged or polar amino acids result in the largest energy change, while the effect is much smaller in the cases of small or hydrophobic residue replacement.

Particular attention deserve the 108 cases of nsSNP mutations that retain the physico-chemical properties of the amino acid present in the target protein-protein complex (case "SAME", e.g. *polar→polar; charged→charged, small→small and hydrophobic→hydrophobic*), and another 159 cases in which the nsSNP mutation changes the physico-chemical properties of the original amino acid (case "DIFF") (Table B-1). In terms of the means of the distributions, there is no significant difference between the "SAME" and "DIFF" distributions. In both cases, the mean of the total binding energy and its electrostatic component are negative quantities, which means that in both of these cases, the target complex is more stable than the nsSNP variant. At the same

time, the mean of the van der Waals energy distribution is positive. The difference is in the standard deviation. The standard deviation of the distribution in "SAME" cases is about two times smaller than the distribution of "DIFF" cases. This indicates that there are cases within the "DIFF" category that result in drastic changes in the binding energy.

Table D.1 Parameters of distribution of total binding energy difference and its components

Group	No	$\Delta\Delta\Delta G_{tot}$		$\Delta\Delta\Delta G_{vdw}$		$\Delta\Delta\Delta G_{el}$	
		mean	std	mean	std	mean	std
All	**264**	-0.86	4.28	-0.05	3.11	-0.78	4.64
OMIM	**45**	-1.65	3.80	-1.03	3.32	-2.35	5.51
Non-OMIM	**219**	-0.70	4.36	0.14	3.03	-0.45	4.39
Polar (P)	**62**	-0.27	3.77	0.38	3.94	-0.83	4.74
P-C	**20**	0.05	5.58	1.98	6.35	-3.00	7.59
P-H	**7**	-1.76	3.16	0.54	0.54	-0.66	1.37
P-P	**28**	-0.29	2.58	-0.50	1.73	0.35	2.07
P-S	**7**	0.37	1.46	-0.75	1.79	0.44	0.85
Charge (C)	**76**	-2.01	6.38	-0.33	2.25	-1.37	6.59
C-C	**25**	-2.16	3.59	-0.45	1.77	-1.76	3.99
C-P	**30**	-1.03	7.52	0.24	1.68	-0.38	8.53
C-H	**3**	-5.84	6.23	1.76	2.01	-1.11	10.86
C-S	**18**	-2.80	6.80	-1.45	0.74	-2.51	5.26
Small (S)	**94**	-0.74	2.39	-0.03	2.49	-0.78	2.58
S-C	**10**	0.56	2.72	2.51	3.58	-4.79	4.39
S-H	**31**	0.04	1.57	0.27	1.83	0.38	1.23
S-P	**20**	-0.88	1.53	0.01	1.37	-0.81	2.16
S-S	**33**	-1.79	2.92	-1.11	2.61	-0.65	1.85
Hydrophobic (H)	**32**	0.32	2.50	-0.36	4.46	0.74	3.23
H-C	**2**	-1.40	0.51	-0.90	0.33	-3.14	0.07
H-H	**19**	0.35	1.72	-0.39	1.70	-0.26	1.13
H-P	**5**	0.56	3.65	2.10	9.45	5.14	5.93
H-S	**6**	0.60	4.01	-2.14	5.61	1.55	1.40
SAME	**108**	-1.09	2.97	-0.66	2.05	-0.57	2.58
DIFF	**159**	-0.72	4.96	0.34	3.60	-0.91	5.61
Conserved	**139**	-0.85	4.94	-0.07	3.71	-1.05	5.55
Non-Conserved	**88**	-0.83	3.51	0.06	2.15	-0.72	3.38
High Conserved	**176**	-0.93	4.84	-0.10	3.45	-1.11	5.16
Low Conserved	**51**	-0.53	2.58	0.27	2.11	-0.26	3.38

Conserved (SI 100%), Non-Conserved (SI 10%-99%), High Conserved (SI 80%-100%), Low Conserved (SI 10%-79%).

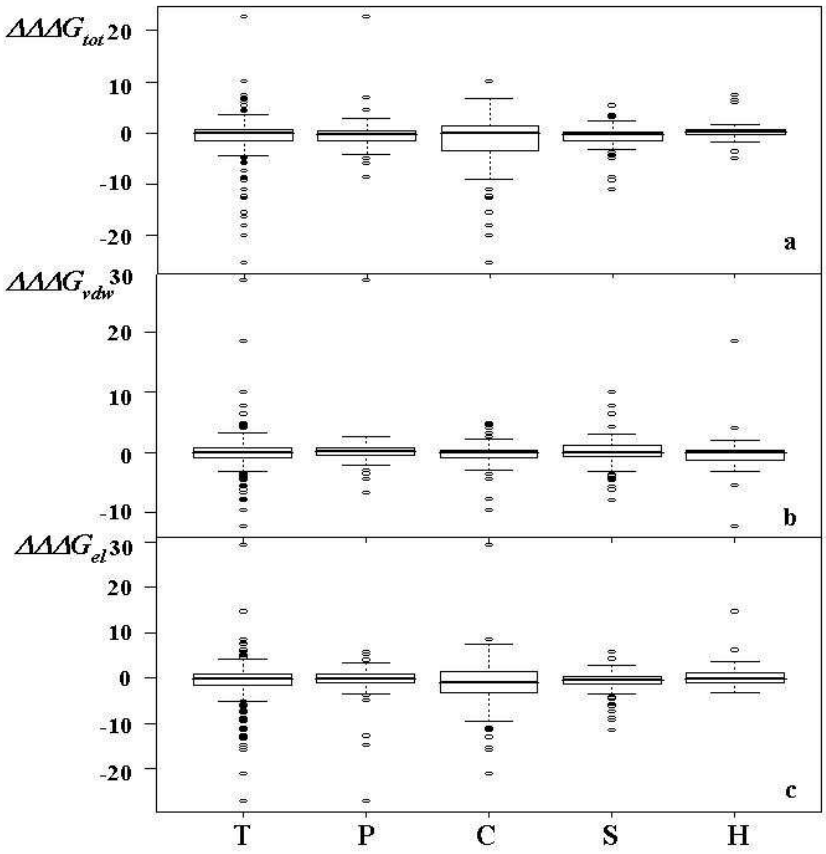

Figure D.1 Distribution of $\Delta\Delta\Delta G_{tot}$*(nsSNP)*, $\Delta\Delta\Delta G_{el}$*(nsSNP)* and $\Delta\Delta\Delta G_{vdw}$*(nsSNP)* in respect with physico-chemical properties of amino acids. T: total, P: polar, C: charged, S: small, H: hydrophobic. The thick black lines show the median, the boxes are drawn between upper and down quartiles and the dotted lines are extended to upper and down whiskers (marked with short horizontal lines).